Beyond Brain Decoding

Methodological and empirical contributions to brain decoding methods and their link to behaviour

Tijl Grootswagers, MSc.

April 2017

Department of Cognitive Science

ARC Centre of Excellence in Cognition and its Disorders

Faculty of Human Sciences

Macquarie University, Sydney, Australia

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Supervisors:

Dr.	Susan G. Wardle	Macquarie University
Professor	Mark A. Williams	Macquarie University
Associate Professor	Thomas A. Carlson	University of Sydney

Table of contents

Summary

Brain decoding methods have transformed the field of Cognitive Neuroscience in the last two decades. As this field has matured, researchers are now asking challenging questions about the interpretation of decoding methods, such as whether the information decoded from neuroimaging data is used by the brain for behaviour. The focus of this thesis is to address the challenge of linking information measured with decoding methods to human behaviour.

In the first empirical chapter, using magneto-encephalography (MEG), I describe a broad set of options for conducting time series decoding studies and test the effects that different options in the decoding analysis pipeline can have on the experimental results. The results show that decisions made at all stages of the analysis can significantly affect the results and interpretation of decoding studies.

In the second empirical chapter, I explore the distance to bound model as a method for linking brain decoding with behaviour. Using MEG decoding, I tested whether this model can account for behavioural changes in reaction time for categorising degraded objects by animacy. I found that the distance to bound model successfully predicted reaction time, accuracy, and decision time parameters derived from a prominent model of decision making. These findings provide evidence for a systematic relationship between decoded brain representations and perceptual decision-making behaviour.

In the third empirical chapter, I examine the distinction between decodable information, and information that can be used in behaviour. Using a searchlight approach on

functional Magnetic Resonance Imaging (fMRI) data, I first investigated where decodable information existed in the brain. Secondly, I assessed where in the brain the decoded information was suitable for “read out” by the brain for behaviour. I found that behaviour can only be predicted from a subset of the locations that had decodable information. These results highlight the distinction between decodable information, and information that is relevant for behaviour in the brain.

In conclusion, this thesis advances current knowledge on brain decoding methods and on approaches to relating brain representations to behaviour, which is a fundamental challenge in cognitive neuroscience. The results show that decodable information has to be interpreted with caution, and emphasize that continuing to develop methods for linking neuroimaging to behaviour is critical for advancing our understanding of the brain.

Author Note

This thesis was prepared in the form of a thesis by publication. The empirical chapters in this thesis present stand-alone research articles. Of these, the first is published, the second is currently in a second round of review, and the third is being prepared for submission. As such, there is a degree of repetition in the chapter introductions, references are given at the end of individual chapters, and the numbering of sections, figures, and tables resets at the beginning of each chapter.

All code for the research presented in this thesis was written in Matlab, some of which has been contributed to the versatile CoSMoMVPA toolbox (http://cosmomvpa.org). The distance to bound searchlight created for chapter four was developed as extension for the CoSMoMVPA toolbox and is currently being prepared for inclusion in the toolbox (https://github.com/Tijl/CoSMoMVPA/tree/decisionvalues). Code for the other chapters is freely available under an open-source license from my online repositories:

Chapter two: https://github.com/Tijl/Grootswagers_etal_2017_decodingtutorial

Chapter three: https://github.com/Tijl/Grootswagers_etal_degraded_objects

Author Statement

I, Tijl Grootswagers, certify that this thesis presents an original piece of research and is written by me. I certify that this thesis has not been previously submitted for a degree, and that it has not been submitted to any university other than Macquarie University, Sydney, Australia. Any sources of information or assistance are appropriately cited and acknowledged. All research presented in this thesis was approved by the Macquarie University Human Research Ethics Committee (reference number: 5201300804, Appendix A).

Signed,

Tijl Grootswagers, April 2017

Acknowledgements

Leaving the Netherlands to pursue a PhD at the other side of the world has turned out to be an excellent decision. Despite of course missing European bread, cheese, and family, the last few years in Sydney have been an amazing adventure. Amongst countless new experiences, it has resulted in this thesis. Of course, this thesis would not have been possible without the support of many others.

First and foremost, I thank my supervisors. Tom, your enthusiasm for science has inspired me greatly. You provided invaluable contributions to my projects at all levels, while at the same time allowing and encouraging me to freely explore. Thank you for supporting my travels to overseas conferences and labs, which has greatly enriched my PhD experience. Susan, thank you for all the great discussions and advice, and especially for the incredibly efficient write-up coaching, which at some points made writing seem almost as fun as programming. Mark, thank you for your support throughout the PhD and for your fast responses to my emails, which were often sent mere minutes before deadlines.

The perception in action research group of the CCD at Macquarie University has been an ideal research environment for me. In particular, I thank Anina Rich, David Kaplan, Matthew Finkbeiner, and the other members of the PARC lab for the weekly discussions, and for providing me with valuable feedback on my research. Many thanks to the CCD administrative team, in particular Lesley McKnight and Teri Roberts, who for me made all administrative procedures much less painful than they were designed to be.

My experience over the last years was greatly enhanced by interactions with other researchers, for which I am very grateful. In particular, I thank: Brendan Ritchie, for introducing me to the lab in my first weeks in Sydney, hosting me in Maryland, and the many insightful discussions. Radoslaw Cichy for hosting me in his lab in Berlin, and for working with me on the final chapter. Andrew Heathcote for his assistance with the reaction time modelling. Dan Butts for letting me spend a week in his lab at the University of Maryland, Hinze Hoogendoorn for hosting me at Utrecht University, and David Leopold and David McMahon for hosting me at NIH. I also thank my MSc supervisors at the University of Nijmegen, Iris van Rooij and Todd Wareham, for setting me on this path.

Thanks to all my fellow students and friends for making this a fun experience. Especially, Lina, Jade, Jon, and Felix for the great times we had in the last years. To my parents, Hans and Christien; our decision to move to Australia was probably the hardest for you, but you have still always supported me in every way possible. Also thanks to my siblings, Sterre, Pol, and Dijk for their encouragement. Finally, Denise; Thank you for your endless support and for accompanying me to the other side of the world, I'm looking forward to our many adventures to come.

Chapter 1

Introduction

In the last two decades, decoding methods have significantly influenced and transformed the field of Cognitive Neuroscience. The application of machine learning algorithms to neuroimaging data allows neuroscientists to "decode" information in the human brain (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Haxby et al., 2001; Haynes, 2015; Haynes & Rees, 2006; Kamitani & Tong, 2005; Kriegeskorte, Goebel, & Bandettini, 2006; Norman, Polyn, Detre, & Haxby, 2006; Tong & Pratte, 2012). Together with increasing computational power, the development and improvement of neuroimaging recording methods and decoding algorithms have made brain decoding highly accessible to scientists. However, a current challenge is to link brain decoding methods to behaviour (de-Wit, Alexander, Ekroll, & Wagemans, 2016; Ritchie, Kaplan, & Klein, in press). This thesis aims to address this challenge by exploring methods that go beyond brain decoding. The first part of the thesis focuses on developing decoding methods for magneto-encephalography (MEG), as they are not as well established as those for functional Magnetic Resonance Imaging (fMRI). The second part focuses on linking fMRI and MEG decoding methods to behaviour.

The brain decoding approach is part of a larger set of analysis tools, known as MultiVariate Pattern Analysis (MVPA) methods. Rather than considering one variable at a time as in traditional neuroimaging analyses, these methods are applied to *patterns* of

activation measured using neuroimaging techniques. Within this framework, decoding refers to the use of machine learning classifiers (see Bishop, 2006) to predict a variable (e.g., what image the participant was looking at) from neuroimaging data, such as functional Magnetic Resonance Imaging (fMRI), electro-encephalography (EEG), or magneto-encephalography (MEG). If the classifier's predictions are better than would be expected by chance, this means that the neuroimaging data contains information about this variable. The ability to decode information from neuroimaging data allows us to non-invasively study information processing in the brain, which improves our understanding of the brain (Carlson et al., 2003; Cox & Savoy, 2003; Haxby et al., 2001; Haynes, 2015; Haynes & Rees, 2006; Kriegeskorte et al., 2006; Norman et al., 2006; O'Toole et al., 2007; Tong & Pratte, 2012). Decoding information from the brain also has direct practical applications, such as in Brain-Computer Interfaces (BCI), which enable paralysed patients to communicate and operate prosthetics (Hatsopoulos & Donoghue, 2009). However, the goals for these applications are to optimally *predict* information from the brain. This differs from the goal of decoding in Cognitive Neuroscience, which is to make *inferences* about the informational content in the brain (Friston et al., 2008; Hebart, Görgen, & Haynes, 2015). The focus of this thesis is on brain decoding methods for Cognitive Neuroscience.

In the field of Cognitive Neuroscience, pattern-based methods were first applied to the domain of visual object recognition (Edelman, Grill-Spector, Kushnir, & Malach, 1998). MVPA approaches have since significantly influenced the field of object recognition (Grill-Spector & Weiner, 2014; Haynes, 2015; Kriegeskorte & Kievit, 2013). The sensitivity of the decoding approach revealed differential brain responses at the single object level (Carlson et al., 2003; Cox & Savoy, 2003; Edelman et al., 1998; Haxby et

al., 2001; Haynes, 2015; Kriegeskorte, Mur, Ruff, et al., 2008; Tong & Pratte, 2012). This allowed testing predictions about the structure of object representations (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, Mur, & Bandettini, 2008; Mur et al., 2013), for example whether they are organized into object categories (e.g., animals, tools, or plants). However, an important question is whether these decoded representations reflect the information that is used by the brain in behaviour. For example, when we categorize an object as an animal, is the brain reading out the category from the same representation that we decode from the neuroimaging data? Answering this question is not possible by decoding the representations alone. In fact, a fundamental challenge in cognitive neuroscience research is to devise ways to address this question (de-Wit et al., 2016; DiCarlo, Zoccolan, & Rust, 2012; Williams, Dang, & Kanwisher, 2007). This thesis aims to contribute to this effort by exploring methods for decoding brain representations and linking them to behaviour.

The main contributions of this thesis will be on brain decoding methods for Cognitive Neuroscience, and methods for linking brain decoding results to behaviour. For the purposes of the thesis, the methods are applied to the domain of visual object recognition and categorization. The reported visual object categorization effects are large, well-described, and replicated multiple times. Moreover, monkey neurophysiology has extensively studied object representations and read-out in primates (Baizer, Ungerleider, & Desimone, 1991; Desimone, Schein, Moran, & Ungerleider, 1985; Felleman & Van Essen, 1991; Freedman, Riesenhuber, Poggio, & Miller, 2001, 2003, Gross, 1973, 1994; Hung, Kreiman, Poggio, & DiCarlo, 2005; Ungerleider, 1982), which can inform human neuroimaging studies about where to look, and what to look for. These aspects make object recognition an excellent platform for exploring how to improve neuroimaging

methods. Thus, while yielding insights into the neural representations of visual objects in the brain, the results presented in this thesis mainly address methods for linking brain and behaviour.

The empirical work of this thesis is organized in three chapters. In chapter two of the thesis, I will focus on the development of MEG decoding methods, as these are currently not well established nor described in the literature. While MEG decoding toolboxes now exist, they are still under active development, and require backgrounds in machine learning and programming. Introducing guidance and establishing protocols is therefore a vital undertaking to ensure a solid foundation for the nascent field of MEG decoding. The chapter presents findings on comparing different parameters in the preprocessing and analysis phases of decoding, which show how they can affect the results and interpretation of decoding studies. Chapters three and four focus on linking these decoding methods to behaviour using the recently proposed distance to bound approach (Ritchie & Carlson, 2016). In chapter three I use MEG decoding to investigate whether a behavioural manipulation that affects reaction times is accompanied by a similar change in distance to bound. In chapter four, I apply the distance to bound approach to fMRI to quantify the dissociation between decodable information and information that can be used by the brain in behaviour. In chapter five, I discuss how the findings from these chapters contribute to the decoding literature in Cognitive Neuroscience, and how they contribute to the challenge of linking decoding methods to behaviour.

In this introductory chapter, I will review the literature, starting with a background section on recent advances in visual object recognition and categorization (section 1). Next, I describe the multivariate pattern analysis approach and its application to neuroimaging

data (section 2). I then review limitations of this approach and discuss how it is challenging to answer the important question of how this information is used by the brain in behaviour (section 3). Finally, I describe a recent approach that links brain decoders to behaviour by modelling the read-out of information (section 4). I conclude the first chapter with an outline of the rest of the thesis (section 5).

1 Visual object recognition and categorization

Humans effortlessly recognize and categorize objects that vary wildly in, for example, size, viewpoint, or lightning (Biederman, 1987; DiCarlo et al., 2012; Logothetis & Sheinberg, 1996; Mahon & Caramazza, 2011; Potter, 1976; Thorpe, Fize, & Marlot, 1996; Ungerleider, 1982). For everyday objects, such as cars, cats, or coffee mugs, the information that falls on the retina is very different when the objects are viewed from another angle, close up or from a distance, in the dark, or partly obstructed. Still, we don't have any problem recognizing cars as cars, and cats as cats, and subsequently accessing semantic information about these categories, for example, that a cat is an animal, apples are edible, and cars are vehicles. The human brain takes about a third of a second to recognize these objects (Grill-Spector & Kanwisher, 2005; Thorpe et al., 1996), which is remarkable for such a difficult problem.

To date, no artificial systems have been created that exactly mimic human performance, but recent developments in artificial deep neural networks have been a giant leap towards creating systems that can perform object recognition and categorization with high accuracies (Cadieu et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012; Sermanet et al., 2013; Simonyan & Zisserman, 2014). However, their workings are in many aspects

different from those in the brain (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Nguyen, Yosinski, & Clune, 2015), and deep neural networks can therefore only provide limited insights into human object recognition (cf. VanRullen, 2017). After decades of research into the computational underpinnings of vision since Marr (1982), the mechanisms and implementation of the brain's efficient solution for the challenges of visual object recognition are not well understood (DiCarlo & Cox, 2007; Grill-Spector & Weiner, 2014) and understanding how the brain solves the challenges in visual object recognition remains a major goal in Cognitive Neuroscience.

To recognize an object, the brain must perform a series of transformations to the image that falls onto our retina (DiCarlo et al., 2012; Grill-Spector & Malach, 2004; Grill-Spector & Weiner, 2014). These involve detecting the edges of an object, segmenting it from the background, and transforming it into abstract representations that are read out when making a perceptual decision about the object or category membership. Behavioural and neuropsychological research has provided a large and important body of knowledge on visual objects (Caramazza & Mahon, 2003, 2006; Caramazza & Shelton, 1998; Logothetis & Sheinberg, 1996). For example, studies on categorization have explored at what level of abstraction (e.g., superordinate 'animal', basic 'bird', and sub-ordinate 'pelican') objects are identified (Grill-Spector & Kanwisher, 2005; Mack & Palmeri, 2010, 2011) and the timing of those stages (Fabre-Thorpe, Richard, & Thorpe, 1998; Kirchner & Thorpe, 2006; Mack & Palmeri, 2011; Thorpe et al., 1996). As the focus of this thesis is on MVPA methods for neuroimaging, a complete review of the literature on object recognition is outside the current scope. In the next sections, I will thus focus on the recent advances on MVPA approaches to neuroimaging data and their applications to understanding the structure of object representations.

1.1 Neural mechanisms of object recognition

The neural underpinnings of how the brain solves the challenges of object recognition have been extensively studied, but are still not well understood. Evidence has put forward the Inferior Temporal (IT) Cortex in macaques and the human homologue Ventral Temporal Cortex (VTC) as a critical structure for object recognition. VTC and IT respond differently to different object categories (Freedman et al., 2003; Grill-Spector & Weiner, 2014; Gross, Rocha-Miranda, & Bender, 1972; Haxby et al., 2001; Hung et al., 2005; Konkle & Caramazza, 2013; Kriegeskorte, Mur, Ruff, et al., 2008; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Zhang et al., 2011), and lesions to these areas result in impaired object processing (Farah, 1990; Konen, Behrmann, Nishimura, & Kastner, 2011). A dominant view of the functioning of the visual object recognition system is that it involves a hierarchy of transformations in the ventral visual stream, embedded in areas between V1 and the VTC (Figure 1). Moving up the hierarchy, areas represent more complex features; from simple visual features to invariant and categorical representations (DiCarlo & Cox, 2007; Hubel & Wiesel, 1962, 1965, Riesenhuber & Poggio, 1999, 2000). The final stage is thought to be a representation that is suitably formatted for read-out by other areas of the brain (DiCarlo & Cox, 2007; Grill-Spector & Weiner, 2014). However, we currently do not know the exact nature of these representations, or how they are used by the brain.

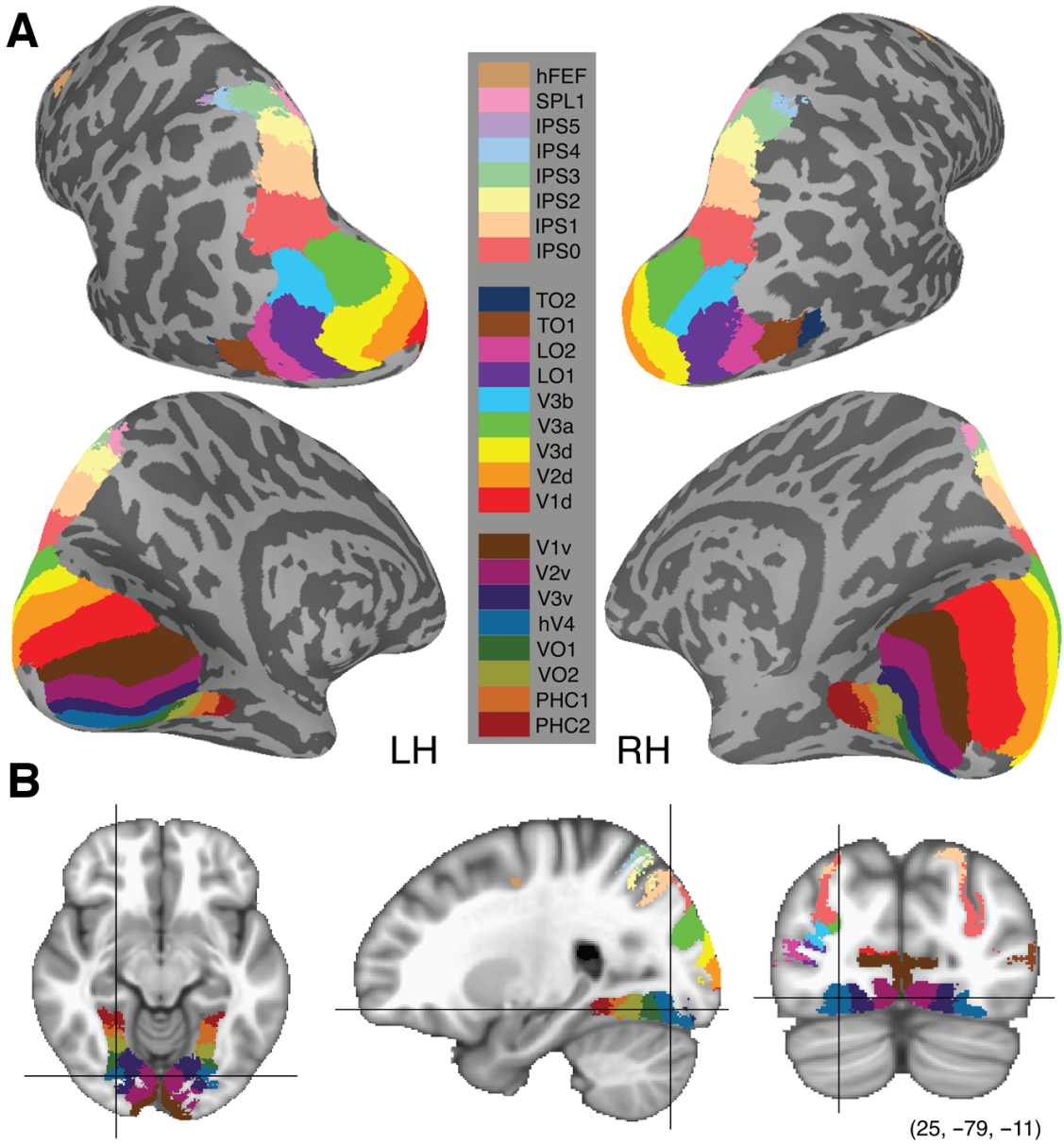

Figure 1. Topological organization of the visual system in the human brain. A. Areas projected onto a surface image of the brain. **B.** The same areas in volume space. Figure adapted from Wang, Mruczek, Arcaro, & Kastner, (2015).

A major advance in visual object research was the application of classifiers to fMRI patterns of activity. MVPA was first applied to visual object recognition studies around 2000 (Carlson et al., 2003; Cox & Savoy, 2003; Edelman et al., 1998; Haxby et al., 2001), showing its potential to distinguish single object exemplars from each other on an almost trial-by-trial basis from patterns of activity in fMRI. The VTC has been a major focus of MVPA studies to date (Carlson, Simmons, Kriegeskorte, & Slevc, 2013; Cox & Savoy, 2003; Grill-Spector & Weiner, 2014; Haxby et al., 2001; Konkle & Caramazza, 2013; Konkle & Oliva, 2012; op de Beeck, Haushofer, & Kanwisher, 2008; Sha et al., 2015). In addition, MVPA studies have explored the representations in the functionally defined object-selective Lateral Occipital Complex (LOC), and neighbouring specialized areas such as the Fusiform Face Area (FFA) have been studied extensively (Downing & Peelen, 2016; op de Beeck, Torfs, & Wagemans, 2008; Peelen, Wiggett, & Downing, 2006). For example, MVPA studies were the first to show that the FFA, an area thought to be a specified module for face processing (Kanwisher, McDermott, & Chun, 1997) contained patterns with more fine grained information about other objects as well (Downing, Chan, Peelen, Dodds, & Kanwisher, 2006; Haxby et al., 2001).

1.2 The animate – inanimate distinction

The major categorical organization principle is thought to be object animacy (Caramazza & Mahon, 2003; Caramazza & Shelton, 1998). Recordings from a set of macaque neurons in IT showed that responses to animate objects were more similar to other animate objects than to inanimate object responses (Hung et al., 2005; Kiani, Esteky, Mirpour, & Tanaka, 2007). In human neuroimaging, animate objects evoke higher activations in the lateral VTC, and inanimate objects in the medial VTC (Chao, Haxby, &

Martin, 1999; Downing et al., 2006; Mahon et al., 2007). Similar conclusions were drawn from MVPA studies, where brain responses to animals were more similar to each other than to inanimate objects and animacy was found to be a highly decodable category (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014; Kriegeskorte, Mur, Ruff, et al., 2008; Proklova, Kaiser, & Peelen, 2016). Figure 2 visualizes the similarity between brain responses and shows how animacy is the major organizational principle. Rather than a pure animacy dichotomy, animacy has been suggested to be organized along a continuous dimension, from more typical animals (e.g., mammals or birds) to less typical, such as bugs or fish (Connolly et al., 2012; Iordan, Greene, Beck, & Fei-Fei, 2016; Konkle & Caramazza, 2013; Sha et al., 2015).

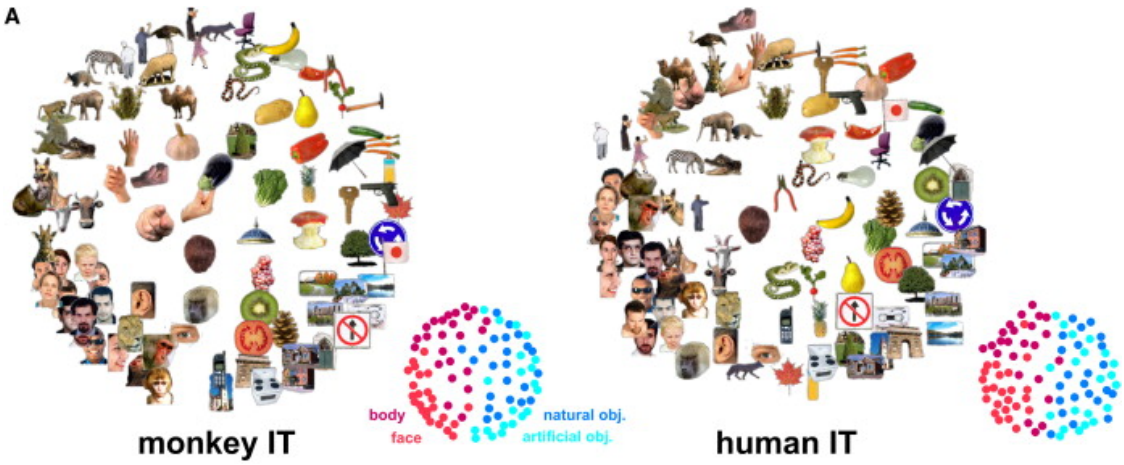

Figure 2. Reconstructed object space from human and macaque IT recordings. Multidimensional Scaling (MDS, see e.g., (Torgerson, 1958)) was used to visualize relative similarities in IT response patterns between different objects. Items close to each other in this space evoke similar brain responses. In this reconstruction, faces cluster together in the bottom left corner, showing that they evoke similar IT responses. In addition, animate objects and inanimate objects are on opposite sides in this space, showing animacy as a major categorical division in human and primate IT. These high level categorical organizations look strikingly similar between the human and monkey reconstructions. Figure from Kriegeskorte, Mur, Ruff, et al., (2008).

1.3 Outstanding questions on the neural representation of objects

The focus of visual object neuroimaging research to date has been on the organization of the structure of object representations in space and time. Divisions such as animacy (Caramazza & Shelton, 1998; Kriegeskorte, Mur, Ruff, et al., 2008; Mahon & Caramazza, 2011; Martin, 2007; Spelke, Phillips, & Woodward, 1995), and object size (Konkle & Caramazza, 2013; Konkle & Oliva, 2012) are well described. The underlying assumption is that these structures are read out by the brain in behaviour (DiCarlo & Cox, 2007; Ritchie, Kaplan, et al., in press). Yet, the relationship between reconstructed neural representations and behaviour is not well described and is likely more complex than currently assumed. For example, different tasks often have no or relatively small effects on the neural representational structures as measured with current neuroimaging methods, but produce completely different behaviour (Harel, Kravitz, & Baker, 2014; Ritchie, Tovar, & Carlson, 2015). In addition, some regions of the brain can have decodable information that is not related to behavioural read-out at all (Williams et al., 2007). A major criticism of MVPA, neuroimaging, and the field of cognitive neuroscience in general is the lack of addressing the relationship between brain and behaviour (Coltheart, 2006; de-Wit et al., 2016; Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Poldrack, 2006, 2010). This criticism is a main focus of the thesis and is expanded on in section 3 of this chapter. I will first describe the MVPA methods in more detail.

2 Multivariate Pattern Analysis (MVPA)

MVPA methods have become standard practice in analysing fMRI data (Cox & Savoy, 2003; Haxby et al., 2001; Haynes, 2015; Norman et al., 2006). The defining commonality of these methods is that they consider the systematic relationship between multiple variables (e.g., voxels in fMRI or channels in MEG). Rather than the univariate approach of treating each voxel as independent and looking for differences in average activation (Friston, Holmes, et al., 1994; Friston, Worsley, Frackowiak, Mazziotta, & Evans, 1994; Worsley, Evans, Marrett, & Neelin, 1992), MVPA methods are used to find differences in patterns of activation in a set of voxels to answer questions about the presence of information (Kriegeskorte & Bandettini, 2007; Kriegeskorte et al., 2006). A popular MVPA approach is to train machine learning classifiers, such as the Support Vector Machine (SVM) to predict the stimulus (or condition) from the data.

2.1 The general MVPA decoding approach

The MVPA approach is illustrated in Figure 3. Patterns of activity are represented as points in a multidimensional space. For simplicity, this is illustrated as two-dimensional space (visualization becomes harder in three or more dimensions, see Section 4.1), but there are as many dimensions as there are features (i.e., fMRI voxels) in the data, and in practice this means dealing with hundreds of dimensions. The data is split up in sets for cross-validation and a classifier is trained to predict the class (e.g., cats vs dogs) of an observation on all-but-one set. The trained classifier is used to predict the class of the items in the left out set. A performance measure is then computed on the predictions, for example the percent of correctly classified items.

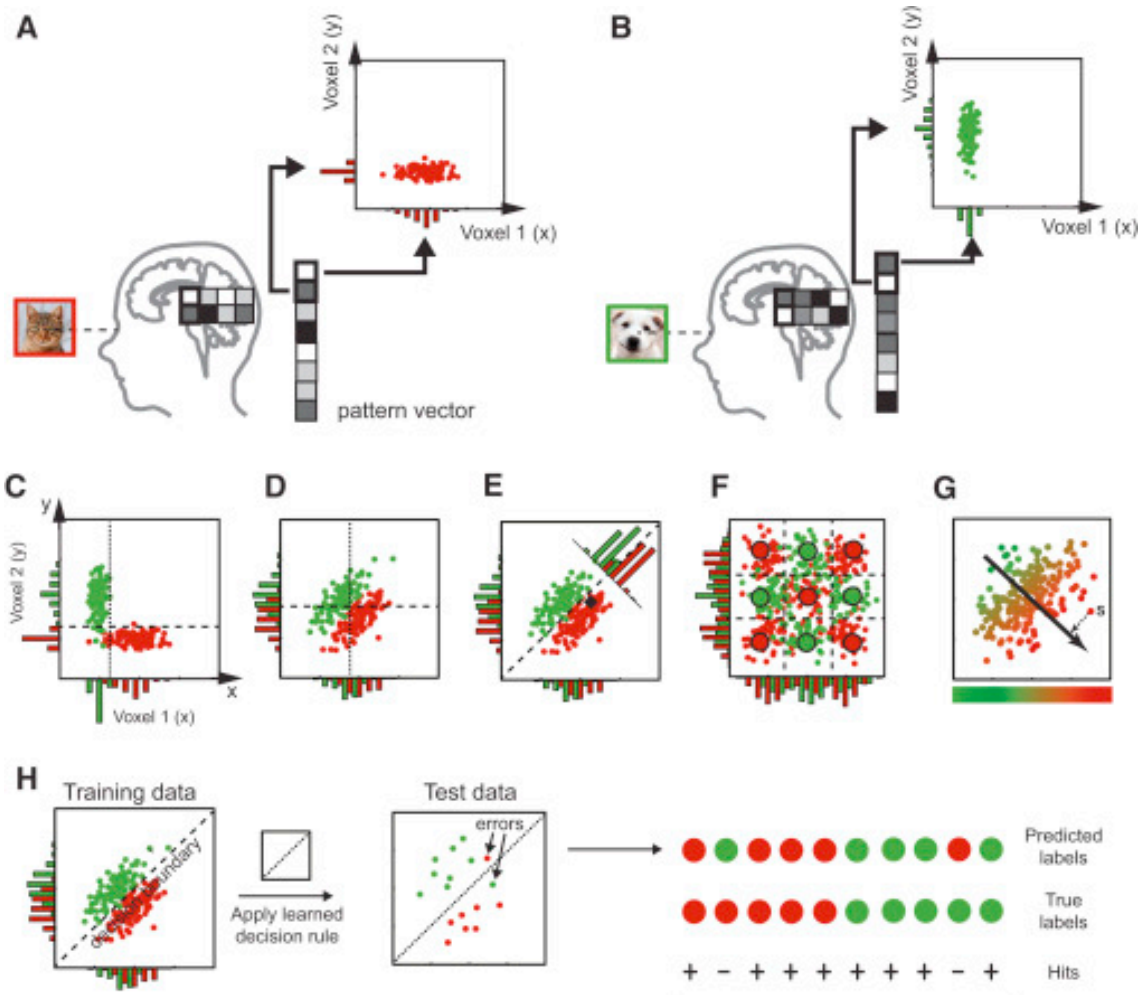

Figure 3. Classification of fMRI patterns of activity. A,B. Activations in response to cat and dog stimuli in 8 voxels are represented as vectors. Points in two-dimensional space represent the activation for two voxels to multiple presentations of cat (red) or dog (green) stimuli. **C.** If the response patterns to the two classes are different within one voxel, they will be visible when analysing the activity in only that voxel (a univariate approach). **D.** When the distributions overlap, the relation between voxels has to be taken into account to discriminate the classes. **E.** Here, a linear decision boundary can be used to discriminate between cat and dog responses. **F.** Here, a linear classifier would not be successful as the classes are not linearly separable. **G.** The classification approach can be seen as a projection onto a one-dimensional decision axis. **H.** Cross-validation involves training the classifier on part of the data (the training set) and testing its discrimination performance on new data (the test set). The classifier's performance can be measured as percent correctly classified items in the test set. Figure adapted from Haynes, (2015).

If the classifier is correct more often than expected by chance, then we can conclude that the signal contains information about these classes, for example the cat responses versus dog responses in Figure 3. The decoding process can then be repeated, for example by performing the decoding on multiple masks of voxels that belong to different brain regions of interest (ROIs; e.g., IT or V1). This allows making inferences on the availability of information per region. Alternatively, a searchlight (Haynes & Rees, 2006; Kriegeskorte et al., 2006; Woolgar, Jackson, & Duncan, 2016) can be used. This involves repeating the decoding at different spatial locations in the brain. At each voxel, a local sphere of neighbouring voxels is extracted and serves as input to the decoding process. The resulting decoding accuracy is then stored at the centre voxel of this sphere, and the process is repeated for all voxels in the brain. This allows mapping the availability of information without the need for a-priori ROI definitions.

Decoding analyses can reveal whether information between conditions is present in the patterns of brain activity. However, we often want to study the underlying representational *structure* of the information, to compare the geometry and transformations of these representations to computational models (Kriegeskorte & Kievit, 2013). Internal object representations have been described and visualized in terms of similarities between their neural or behavioural responses (Edelman, 1998; Edelman & Duvdevani-Bar, 1997; Edelman et al., 1998; op de Beeck, Torfs, et al., 2008; op de Beeck, Wagemans, & Vogels, 2001). Expressing internal representations as similarities inspired the development of Representational Similarity Analysis (RSA) (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Mur, Bandettini, & Kriegeskorte, 2009). This technique has facilitated testing models of the high level categorical organization of human IT (Carlson, Simmons, et al., 2013; Connolly et al., 2012; Haxby,

Connolly, & Guntupalli, 2014; Kriegeskorte, Mur, Ruff, et al., 2008; Mur et al., 2013; Sha et al., 2015). Moreover, RSA has been used to compare the representational structure of information between species and modalities (Cichy et al., 2014; Cichy, Pantazis, & Oliva, 2016; Kriegeskorte, Mur, & Bandettini, 2008; Mur et al., 2013). For example, similar categorical structures with animacy as the major organizational principle were found in the organization of IT in both humans and macaques (Kriegeskorte, Mur, Ruff, et al., 2008).

2.2. Decoding in fMRI and MEG

The vast majority of MVPA studies use fMRI, which has therefore been the focus of the development of MVPA methods. This has resulted in a large body of reviews and tutorials (Cox & Savoy, 2003; Formisano, De Martino, & Valente, 2008; Haxby et al., 2014; Haynes, 2015; Haynes & Rees, 2006; Mur et al., 2009; Norman et al., 2006; Pereira, Mitchell, & Botvinick, 2009; Schwarzkopf & Rees, 2011; Tong & Pratte, 2012). In addition, several toolboxes have been developed for performing MVPA on fMRI (Hanke et al., 2009, 2009; Hebart et al., 2015; Oosterhof, Connolly, & Haxby, 2016) making the methods increasingly more available to researchers. The popularity of fMRI is due to it having the highest spatial resolution of non-invasive methods. The most common 3-Tesla scanners record the hemodynamic response in 1-2mm^3 voxels. However, the hemodynamic response is a slow process, and fMRI only samples the signal roughly once every second. Therefore, fMRI has a very low temporal resolution. In contrast, the speed of processing in the brain is in the order of milliseconds, for example the time it takes for visual information to arrive in V1 is around 60 to 70 milliseconds (Nowak & Bullier, 1997; Thorpe et al., 1996). Because of this, fMRI is less suited for studying

temporal dynamics. Instead, the same approaches can be applied to MEG (or EEG), which is generally sampled at 1000Hz, and therefore allow millisecond by millisecond analysis of brain processes.

Recently, applying MVPA methods on MEG data has become increasingly popular. Most applications to date of decoding in MEG have been in the visual domain (Contini, Wardle, & Carlson, in press). For example, examining the time course of emerging object and category representations (Barragan-Jason, Cauchoix, & Barbeau, 2015; Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011; Carlson, Simmons, et al., 2013; Cauchoix, Crouzet, Fize, & Serre, 2016; Isik, Meyers, Leibo, & Poggio, 2014; Kaiser, Azzalini, & Peelen, 2016; Simanova, van Gerven, Oostenveld, & Hagoort, 2010), orientation decoding (Cichy, Ramirez, & Pantazis, 2015; Ramkumar, Jas, Pannasch, Hari, & Parkkonen, 2013; Wardle, Kriegeskorte, Grootswagers, Khaligh-Razavi, & Carlson, 2016), and visual scenes (Cichy, Khosla, Pantazis, & Oliva, in press; Kaiser, Oosterhof, & Peelen, 2016). MEG decoding has also been combined with fMRI to compare the decodable information between the two, and to get a full picture of the temporal and spatial dynamics of objects in the brain (Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Kaiser, Azzalini, et al., 2016). Furthermore, it allows comparing classifier performance over time (King & Dehaene, 2014b), which has been used in the domain of object decoding to show that emerging object representations are not static, but highly dynamic (Carlson, Tovar, et al., 2013; Cichy et al., 2014; Contini et al., in press; Isik et al., 2014; Kaiser, Azzalini, et al., 2016; Kaiser, Oosterhof, et al., 2016). Taken together, the limited number of MEG decoding studies have already produced exciting results, and show a rich potential for its use in future cognitive neuroscience research. Yet, MEG decoding methods are not as extensively documented as those for fMRI. In contrast to fMRI, there

are currently no reviews or tutorials that introduce the techniques and considerations that play a role in time-series decoding. To address this gap, chapter two of this thesis compares the effects of different analysis options for time-series decoding and is intended to serve as a guideline to inform future time-series decoding studies.

3 Interpreting MVPA results

MVPA encompasses a set of extremely powerful and sensitive tools for neuroimaging data analysis. However, as Cox and Savoy noted in one of the first MVPA decoding studies, the observation that information can be extracted from the neuroimaging data does not imply that the brain is using this information (Cox & Savoy, 2003). While this is true for any correlational analysis, it is not always made explicit in interpreting MVPA studies. Finding decodable information is often communicated as being equivalent to identifying the underlying neural code or representations, even though neuroimaging researchers are in general aware of the limits of the correlational analysis (de-Wit et al., 2016). de-Wit et al argue for a shift in perspective, and to focus on searching for decodable information that can be shown to be read-out by the brain. Next, I will discuss three examples that argue for more caution in interpreting decodable information; First, we don't have a complete description of whether classifiers are able to decode information that is organized at a smaller resolution than we measure with fMRI and MEG (3.1). Secondly, it is not trivial to determine and disentangle the sources of decodable information that are confounds from those that contribute to the organization of information in the brain (3.2). Third, finding decodable information in neuroimaging data does not imply that this information is available to the brain as well (3.3). Finally, I

review proposed methods that address these issues by working towards linking decoded information to behaviour (3.4).

3.1 An example from orientation decoding

Early applications of MVPA demonstrated that the orientation of a stimulus can be decoded from V1 (Haynes & Rees, 2005; Kamitani & Tong, 2005). This was surprising, considering that V1 neurons are organized by their preferred orientations in columns at a much smaller scale than the resolution of standard fMRI data (Boynton, 2005; De Martino et al., in press; Hubel & Wiesel, 1974; Hubel, Wiesel, & Stryker, 1978; Yacoub, Harel, & Uğurbil, 2008). The authors argued their results were due to hyperacuity: the classifiers were picking up variability in orientation preference at the single voxel level (Kamitani & Tong, 2005). The hyperacuity account assumes that because voxels randomly sample the orientation-selective neurons, each voxel will have a small bias in the proportions of orientation selective neurons. This sparked a debate about the source of orientation decoding, centred around whether large scale, systematic cortical orientation biases were contributing to the decoding (Freeman, Brouwer, Heeger, & Merriam, 2011; Freeman, Heeger, & Merriam, 2013; Mannion, McDonald, & Clifford, 2009), an alternative explanation to the hyperacuity account. To date, after decades of successful orientation decoding, it is still debated whether small scale biases contribute to the decoding (Alink, Krugliak, Walther, & Kriegeskorte, 2013; Carlson & Wardle, 2015; op de Beeck, 2010; Pratte, Sy, Swisher, & Tong, 2016; Wardle, Ritchie, Seymour, & Carlson, 2017). This debate illustrates that the underlying source of the decodable information is not always known (Carlson & Wardle, 2015; Naselaris & Kay, 2015; Tong & Pratte, 2012).

3.2 Interpreting category decoding

It is not trivial to make inferences on the source of information that is used for decoding, and it is equally hard to control for it. Consider the source of decodable category information, as described in Section 1.1. In MEG decoding of categories and exemplars, onsets for category and exemplar decoding are reported very early in the time series. For example, several studies have found onsets of decodable exemplar information and category information starting around 60-70ms (Carlson, Tovar, et al., 2013; Cauchoix et al., 2016; Cichy et al., 2014). Even though the peak decoding time points in these studies were much later, the onset of decoding indicates the earliest availability of any information in the signal. These early results suggest that they are based on low level visual responses, as information does not reach IT until around 100ms (Thorpe et al., 1996). Moreover, object categories are known to co-vary with visual features (Bar, 2003; Gaspar & Rousselet, 2009; Honey, Kirchner, & VanRullen, 2008; Long, Konkle, Cohen, & Alvarez, 2016; Vanrullen, 2011). Such features are also represented in neural patterns at higher level areas of processing (Andrews, Watson, Rice, & Hartley, 2015; Proklova et al., 2016; Rice, Watson, Hartley, & Andrews, 2014) and classifiers can exploit these features (Vanrullen, 2011). If a classifier can successfully use the low-level responses to extract category information, the brain might be doing the same (Crouzet & Thorpe, 2011). It is therefore not trivial to separate true category decoding effects from confounding or covarying features (Carlson & Wardle, 2015; Grill-Spector & Weiner, 2014; Kaiser, Azzalini, et al., 2016; Mahon & Caramazza, 2011; op de Beeck, Haushofer, et al., 2008; Proklova et al., 2016; Tong & Pratte, 2012; Wardle & Ritchie, 2014).

3.3 Are decoding studies measuring brain processes?

The ambiguity about the source of information (as mentioned in the previous sections) means that the interpretation of decoding studies in terms of internal representations is not straightforward. If there are confounds in the design or stimuli that could be picked up by the classifier, it must be shown that MVPA is extracting the condition of interest and not those confounds (Ramsey et al., 2010; Ritchie, Bracci, & op de Beeck, in press; Todd, Nystrom, & Cohen, 2013; Woolgar, Golland, & Bode, 2014). And even with all possible confounds controlled for, if one brain area shows significant decoding, but another area does not, does this mean information is only represented in the former? Could it be possible that the information that is measured with MVPA is a by-product of a different underlying brain process? Might the brain use a completely different decoding mechanism?

These questions speak to a general criticism and limitation of decoding studies: An implicit assumption in MVPA studies is that if information can be decoded from a brain region, then this information is explicitly represented in this region and used by the brain in behaviour (Haynes & Rees, 2006; King & Dehaene, 2014a; Kriegeskorte & Kievit, 2013; Misaki, Kim, Bandettini, & Kriegeskorte, 2010). For example, in one of the early decoding studies, Haynes and Rees, (2006) state that individual introspective mental events can be recovered from brain activity when the underlying neural representations are decodable, suggesting that finding a decodable signal implies recovering an internal mental state. However, decodability alone is not enough evidence to make claims about representations in the brain and their use in behaviour (Cox & Savoy, 2003; de-Wit et al., 2016; Forstmann & Wagenmakers, 2015; Klein, 2010; Krakauer et al., 2017;

Poldrack, 2006; Ritchie, Kaplan, et al., in press). For instance, in one of the first reviews of the decoding literature, Cox and Savoy (2003) argue that decoding information does not imply that this information is used by the brain, and that experimenters must apply caution with interpreting the nature of the decodable signal. In a recent thought-provoking article, de-Wit et al., (2016) argue that in order to measure information in the brain, one must show that the information is available to the brain, rather than available to the experimenter, by showing that the brain uses the information (de-Wit et al., 2016).

3.4 Linking decoded representations to behaviour

To address the problems described above, several studies have gone beyond the decoding of information in order to find information that relates to behaviour. An advantage of using classifiers for decoding is that their performance can be compared with behaviour (Naselaris, Kay, Nishimoto, & Gallant, 2011). If human performance on the same stimuli can be predicted from the classifier performance, there is evidence for a relation between the decoding and behaviour (Raizada, Tsao, Liu, & Kuhl, 2010; van Bergen, Ji Ma, Pratte, & Jehee, 2015; Walther, Caddigan, Fei-Fei, & Beck, 2009; Williams et al., 2007). For example, Williams et al., (2007) showed that classifier performance in LOC was predictive of task performance and that performance in early visual cortex was not, suggesting that the information in LOC was related to behaviour. Philiastides & Sajda, (2006) varied stimulus coherence in a signal detection task and found that classifier performance matched human psychometric functions. Another approach to linking decoded neural spaces to behaviour is using RSA, where a neural representational dissimilarity matrix (RDM) can be compared to a behavioural dissimilarity matrix. For example by directly comparing them to RDMs of human similarity

ratings (Bracci & op de Beeck, 2016; Mur et al., 2013; Redcay & Carlson, 2015; Wardle et al., 2016), or RDMs that reflect human performance (Cohen, Alvarez, Nakayama, & Konkle, 2016; Proklova et al., 2016). A recently proposed method to link neural decoders to behaviour is the distance to bound approach (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Ritchie & Carlson, 2016; Ritchie et al., 2015), which treats linear neural decoding classifiers as an observer under signal detection theory (Green & Swets, 1966) to predict behaviour. Chapters 3 and 4 of thesis build upon the neural distance to bound approach, which is described in detail in the next section.

4 The neural distance to bound approach

The classifiers used in brain decoding studies show whether information about the experimental condition is discriminable in the multidimensional patterns of activation. However, as described in the previous section, they do not show directly what this information is, or whether this information is being used by the observer (de-Wit et al., 2016; Ritchie, Kaplan, et al., in press; Williams et al., 2007). One aim of this thesis is to address this difficulty by studying the recently proposed neural distance to bound approach (Ritchie & Carlson, 2016). In summary, this approach entails computing the difficulty of individual exemplars for the classifier that was used to decode the brain activation patterns. These difficulties are then compared to a behavioural measure of human difficulty (e.g., reaction times) for the same exemplars on the same classification task. If the behavioural difficulty can be predicted from the classifier difficulty, this shows that the brain could be using the same information as the classifier.

Linear classifier models are a popular choice in decoding studies partly because of their biological plausibility. They can be viewed as a single neuron that reads out information from the input patterns (DiCarlo & Cox, 2007; Misaki et al., 2010). In addition, the classifier uses a decision boundary to separate the multidimensional space into two categories. This attribute closely matches with models of human decision processes under signal detection theory (Ashby, 2000; Ashby & Maddox, 1994; Green & Swets, 1966). Building on this observation, Ritchie & Carlson, (2016) propose that evidence for the classifier's decision can serve as evidence in models of observer decision behaviour. If the brain is using the same neural activation space for a decision, then the evidence for the neural decoder will predict observer categorization difficulty (Carlson et al., 2014; Ritchie & Carlson, 2016). The current section explains this approach in more detail, starting with a measure of evidence for neural decoders (4.1). I then describe how evidence plays a major role in current models of human decision behaviour and how the distance to bound approach links these two concepts (4.2). Finally, the two studies that implemented the approach to date are summarized (4.3) and I discuss the outstanding questions of this approach that this thesis aims to address (4.4).

4.1 A neural hyperplane for classification

The machine learning classifiers used in MVPA are relatively simple linear models. Whether one uses support vector machines (SVM), linear discriminant analysis (LDA), or Gaussian naïve Bayes (GNB), they all tend to perform similarly (Misaki et al., 2010; Pereira et al., 2009). One thing these classifiers have in common is that they all compute a decision function of the form *predicted*(x) = Class0 *if* w x + $c < 0$ *else* Class1. That is, during the training phase, classifiers maximize separation of two classes by computing

a weight vector *w* and a constant *c*. One weight is assigned to each feature. Then, predictions are made by multiplying feature values with the weight vector, and testing whether the resulting value is larger or smaller than zero. Note that $w \times + c = 0$ is a decision hyperplane in multidimensional feature space (Figure 4): items on one side of this hyperplane are predicted as one class, and vice versa. Some items lie closer in multidimensional space to this hyperplane than others (Figure 4). Their decision values are smaller (closer to zero), while others are further away and have larger decision values. For prediction, only the sign of the decision value matters, positive for Class1 and negative for the other. However, the decision values also reflect a measure of confidence for the classifier. To illustrate, if we gather a new set of data for all our items and predict their classes, the items will likely shift due to noise in the measurement. Items with small decision values (close to the hyperplane), could then end up being predicted incorrectly as the other class. Items further from the hyperplane have a higher chance to be correctly predicted in repeats than those closer to the hyperplane. Thus, using this distance property, for every item in our stimulus set, we can obtain a confidence score of the item belonging to class X or Y. Next, I will discuss how this score can be directly linked to models of (human) decision making.

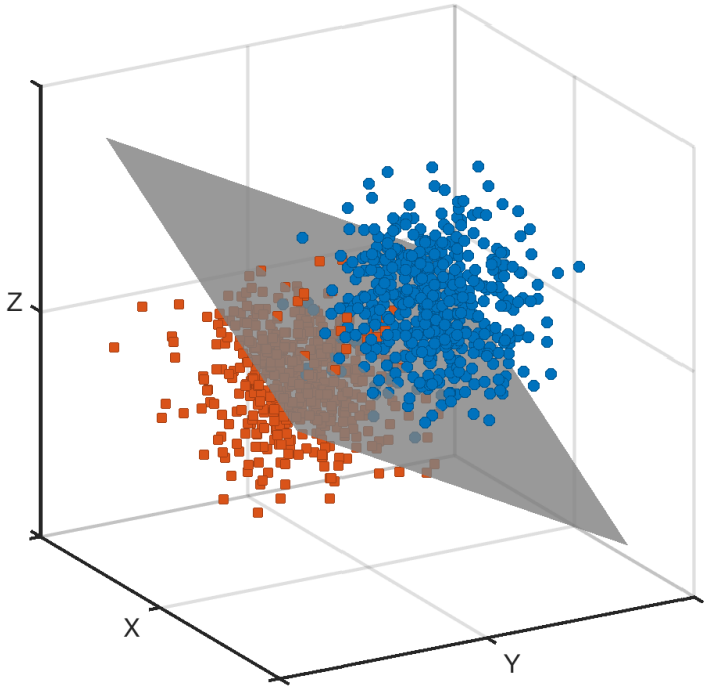

Figure 4. Three-dimensional space with a separating hyperplane. Simulated data. Points in this space represent for example measured activation in three fMRI voxels or three MEG channels. The grey plane represents a hyperplane that separates the three-dimensional space into two. In a decoding setting, items on one side of the hyperplane would be classified as blue circles, and items on the other side as red squares. Here, some items would be misclassified. For each item, a distance to the hyperplane can be computed.

4.2 Evidence accumulation with distance to the hyperplane

The decision hyperplane used in classification can be thought of as a decision boundary in classic signal detection theory (Green & Swets, 1966). Signal detection theory specifies the relation between the stimulus and behaviour. The input for a decision is sampled from distributions on an evidence dimension, where the ratio between evidence determines the response, leading to a decision boundary on the evidence dimension (Figure 5A). The closer an item is to the boundary, the less evidence there is for the response. The amount of available evidence has a major influence on human observer

decision behaviour. Decision making models parameterize human decision making behaviour. The most popular models are evidence accumulation models. These models in general represent each choice as a separate accumulator that gathers evidence over time (e.g., Brown & Heathcote, 2008; Gold & Shadlen, 2007; Smith & Ratcliff, 2004; Stone, 1960). Once one accumulator's evidence reaches a certain threshold, the decision is made. The choice accumulators can vary in, for example, their starting position, accumulation speed, decision threshold. By varying these parameters, evidence accumulation models have been shown to be able to predict behaviour for a variety of tasks. For instance, a response bias to one choice can be accounted for by any combination of lower decision threshold, higher starting position or faster drift rates.

Consider a trivial instantiation of an accumulator model for a binary categorization task. The accumulators for both choices start at the same point (e.g., we assume no priors or variation in starting points), and their thresholds are the same too (e.g., no bias). The only difference in parameters that can account for differences between the two choices is drift rate (the speed of evidence accumulation). At each iteration of the model, the accumulators receive another unit of evidence. If one accumulator reaches a threshold, then the response for that accumulator 'wins' (i.e., the decision is made). Thus, the accumulation speed depends purely on the strength of the evidence, and determines the time of the decision. The decision time is behaviourally measured as reaction time. Now assume that our neural multidimensional space (e.g., fMRI voxel activations) is an accurate representation of the brain's internal space that the observer uses as input for a categorization decision. Our classifier's hyperplane in this space can then be used as the decision boundary for categorization, and the classifier's confidence for a decision (i.e., distance to the hyperplane) serves as evidence for the observer's decision. This

then manifests in accumulation rates, and reaction times for the decision (Figure 5). Thus, if the brain uses the neural multidimensional space for a decision, then the distance to the classifier boundary in this space negatively correlates with reaction times for this decision (Carlson et al., 2014; Ritchie & Carlson, 2016).

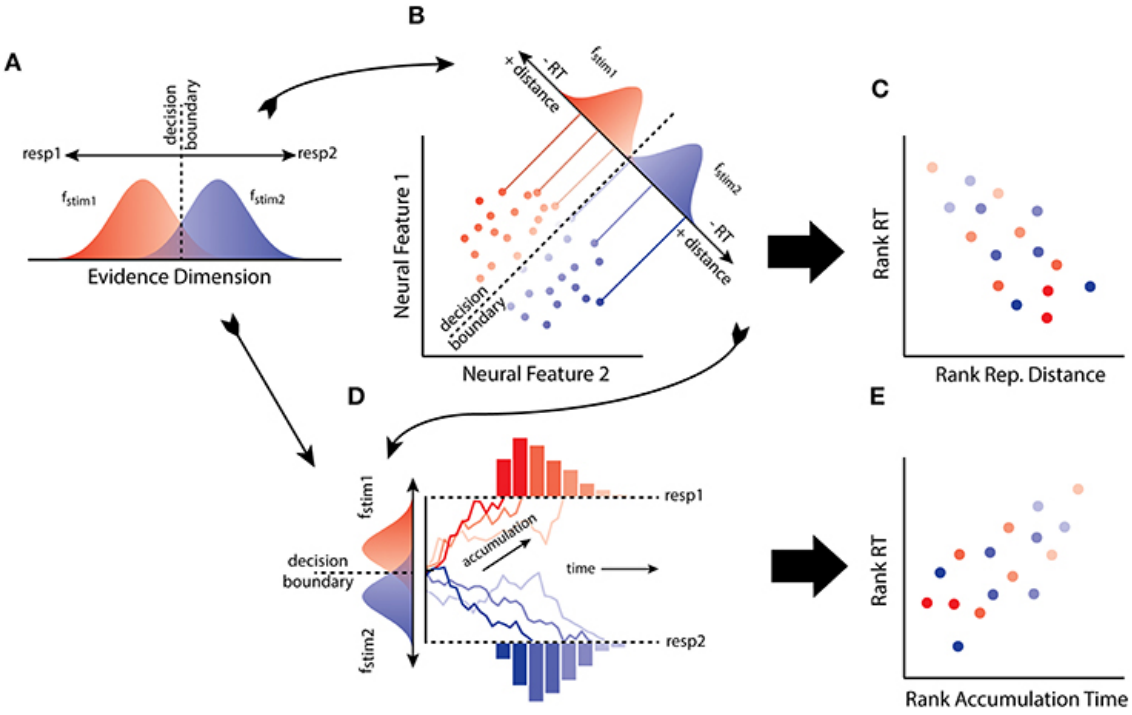

Figure 5. Distance to classifier boundary approach. A. In signal detection theory, the distance from a decision boundary defines evidence for a decision. **B.** A similar decision boundary can be obtained from a neural decoding classifier. **C.** The prediction of the distance to bound approach is that RT decreases as a function of distance to the classifier boundary. **D.** Illustration of how the distance to boundary can be used as accumulation rate in an evidence accumulation model. **E.** Using the neural distance to boundary as drift rate in an evidence accumulation model predicts a positive correlation between accumulation time and RT. Figure from Ritchie & Carlson, (2016).

4.3 Predicting observer reaction times from neural decoders

The hypothesis that distance to a classifier hyperplane predicts RT was tested on representational structures obtained from Regions of Interest (ROIs) of IT and early visual cortex (Carlson et al., 2014). Observer RTs for object animacy were correlated to distances to an animacy classifier boundary on the fMRI ROIs. Their findings showed that the IT ROI had a strong correlation between distance to the boundary and reaction time. Correlations for the early visual cortex were also present but much lower and less reliable. This was taken as evidence to support the involvement of IT in categorization tasks. Importantly, the fMRI data were independent from the RT data, and subjects in the MRI scanner did not have an object-related task, thus the correlations could not be explained by, for example, subject-specific differences in attention. In a follow-up study, the time course of these correlations was explored using MEG decoding (Ritchie et al., 2015), and the same negative relationship between neural distance from the boundary and RT was observed. Moreover, the correlations did not change when subjects were actively performing an animacy task, compared to an orthogonal task in the scanner (Ritchie et al., 2015). These results showed that the distance to bound approach is a step towards linking brain spaces directly to behaviour, addressing a fundamental problem in cognitive neuroscience (de-Wit et al., 2016; Forstmann & Wagenmakers, 2015; Klein, 2010, 2016; Poldrack, 2006; Ritchie, Kaplan, et al., in press). However, there are a number of outstanding questions that need to be addressed in order to fully assess the potential of the distance to bound approach.

4.4 Outstanding questions

Previous distance to bound research found correlations with representational distance obtained from fMRI ROIs (Carlson et al., 2014). These results showed that correlations between distance to the classifier boundary and reaction time in early visual cortex were smaller than the correlations in IT, which was taken as evidence for IT's involvement in categorical decision making. However, some non-zero correlations were found in Early Visual Cortex (EVC). Because the ROIs were relatively large and included multiple visual areas, it is possible that only some parts had correlations with reaction time. A more spatially fine-grained analysis could reveal the full organization of the distinction between decodable category information, and where this information can be read-out in behaviour. Chapter four explores this by combining the distance to bound approach with an fMRI searchlight analysis to localize and quantify the distinction between decodable information and information that can be read-out in behaviour.

Previous studies have found correlations between distance to boundary and reaction time in both fMRI and MEG, regardless of whether participants were actively categorizing the stimuli, or performing an orthogonal task in the scanner (Carlson et al., 2014; Ritchie et al., 2015). While this is evidence that the brain can use the decodable category information, more evidence would come from showing a systematic relationship. For example, showing that manipulating one variable (e.g., reaction time) predicts a change in the other (e.g., distance to boundary) would provide strong evidence in favour of a systematic relationship between distance to boundary and reaction time (cf. Klein, 2016). This systematic relationship was explored in chapter three: The categorization task was

made more difficult, and we tested whether this manipulation of categorization reaction times resulted in a corresponding reduction in distance to boundary.

A general limitation of both studies using the neural distance to bound approach to date is that they used the same stimuli (Carlson et al., 2014; Ritchie et al., 2015). This stimulus set has been used numerous times in visual object research to date (e.g., Carlson, Tovar, et al., 2013; Cichy et al., 2014; Kiani et al., 2007; Kriegeskorte, Mur, Ruff, et al., 2008). Because animacy has been reliably shown as a categorical organization of these stimuli (see Figure 2), it is a good platform for testing novel methods, such as the distance to bound approach. However, it is therefore not known whether the distance to bound approach generalizes to other stimuli and other tasks. For instance, the stimuli that were used included human faces (Carlson et al., 2014; Ritchie et al., 2015), which have generally fast RTs (Crouzet, Kirchner, & Thorpe, 2010). The same stimulus set was used in other studies which showed that the faces in this specific set were also highly decodable (Carlson, Tovar, et al., 2013; Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Kriegeskorte, Mur, Ruff, et al., 2008). Thus, face stimuli could drive the correlation with RT. This raises the question of whether a correlation between distance and reaction time is also possible in data that does not contain human face stimuli, which I investigate in chapter three of this thesis.

Another observation in the distance to bound studies to date was that the animate exemplars were driving the correlations with RT (Carlson et al., 2014; Ritchie et al., 2015). When restricting the correlation to inanimate exemplars, little to no correlations were found. The lack of correlations for inanimate exemplars was argued to be caused by the negatively defined category, i.e., inanimate is defined as ‘not animate’ (Carlson et

al., 2014). This would predict that such an asymmetry does not emerge when using two positively defined categories (e.g., categorizing humans vs animals), a prediction tested in chapter four of this thesis. However, as stated earlier, if human faces are the source of these correlations, then this could explain why correlations are limited to animate stimuli. Taken together, the generalizability of the distance to bound approach, and whether it solely depends on human face stimuli are critical questions that require further investigation. Therefore, in chapters 3 and 4 of this thesis, I examine specific predictions of the distance to bound approach and test its robustness and generalizability to other stimuli and tasks.

5 Overview of thesis

In chapter two, I explore methods for decoding brain representations from time-series neuroimaging data (e.g., MEG). Although the MEG and fMRI decoding approaches are conceptually similar, there are important distinctions, and there are no existing standards for decoding time-varying brain activity. Therefore, I first describe a broad set of options for conducting time series decoding studies. Secondly, using MEG data, I tested the effects that different options in the decoding analysis pipeline can have on the experimental results. The results showed how decisions made at all stages of the analysis can significantly affect the results and interpretation of decoding studies.

In the third chapter of the thesis, I build on the distance to bound approach by testing specific predictions of the approach. Using MEG decoding, I tested whether the neural distance to bound approach could account for increased difficulty in categorizing visually degraded stimuli. Secondly, I tested whether, in addition to reaction times, distance to

bound predicts drift rates as estimated by an accumulator model of decision making. The results showed that distance to the classifier boundary successfully predicted reaction time, accuracy, and drift rates for categorizing the degraded visual objects.

In chapter four, I further examine the relationship between decodable information and behaviour by creating spatially unbiased maps of where the neural readout model could be used to predict behaviour. Using a searchlight approach on fMRI data, I first test for decodable object category information in local voxel clusters, and second, assessed whether this information can also be used to predict categorization reaction times. The results show that decodable information exists along the entire ventral and dorsal visual streams, but that behaviour can only be predicted from a subset of those locations, mostly in the anterior ventral visual stream. These results show a distinction between decodable information, and information that is more likely to be used by the brain behaviour.

In chapter five, I summarize all findings and discuss how they contribute to the current knowledge on brain decoding methods, and to relating brain representations to behaviour. I will describe limitations of the approaches taken in this thesis, discuss outstanding questions, challenges, and give future directions to the field before drawing general conclusions based on this work.

6 References

Alink, A., Krugliak, A., Walther, A., & Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Frontiers in Psychology*, *4*, 493. https://doi.org/10.3389/fpsyg.2013.00493

Andrews, T. J., Watson, D. M., Rice, G. E., & Hartley, T. (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *Journal of Vision*, *15*(7), 3–3. https://doi.org/10.1167/15.7.3

Ashby, F. G. (2000). A Stochastic Version of General Recognition Theory. *Journal of Mathematical Psychology*, *44*(2), 310–329. https://doi.org/10.1006/jmps.1998.1249

Ashby, F. G., & Maddox, W. T. (1994). A Response Time Theory of Separability and Integrality in Speeded Classification. *Journal of Mathematical Psychology*, *38*(4), 423–466. https://doi.org/10.1006/jmps.1994.1032

Baizer, J. S., Ungerleider, L. G., & Desimone, R. (1991). Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. *Journal of Neuroscience*, *11*(1), 168–190.

Bar, M. (2003). A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609. https://doi.org/10.1162/089892903321662976

Barragan-Jason, G., Cauchoix, M., & Barbeau, E. J. (2015). The neural speed of familiar face recognition. *Neuropsychologia*, *75*, 390–401. https://doi.org/10.1016/j.neuropsychologia.2015.06.017

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4). springer New York.

Boynton, G. M. (2005). Imaging orientation selectivity: decoding conscious perception in V1. *Nature Neuroscience*, *8*(5), 541–542. https://doi.org/10.1038/nn0505-541

Bracci, S., & op de Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. *Journal of Neuroscience*, *36*(2), 432–444. https://doi.org/10.1523/JNEUROSCI.2314-15.2016

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7*(8), 354–361. https://doi.org/10.1016/S1364-6613(03)00159-1

Caramazza, A., & Mahon, B. Z. (2006). The organisation of conceptual knowledge in the brain: The future's past and some future directions. *Cognitive Neuropsychology*, *23*(1), 13–38. https://doi.org/10.1080/02643290542000021

Caramazza, A., & Shelton, J. R. (1998). Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience*, *10*(1), 1–34. https://doi.org/10.1162/089892998563752

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, *11*(10), 9. https://doi.org/10.1167/11.10.9

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction Time for Object Categorization Is Predicted by Representational Distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142. https://doi.org/10.1162/jocn_a_00476

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of Activity in the Categorical Representations of Objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. https://doi.org/10.1162/jocn.2003.15.5.704

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2013). The Emergence of Semantic Meaning in the Ventral Temporal Pathway. *Journal of Cognitive Neuroscience*, *26*(1), 120–131. https://doi.org/10.1162/jocn_a_00458

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1. https://doi.org/10.1167/13.10.1

Carlson, T. A., & Wardle, S. G. (2015). Sensible decoding. *NeuroImage*, *110*, 217–218. https://doi.org/10.1016/j.neuroimage.2015.02.009

Cauchoix, M., Crouzet, S. M., Fize, D., & Serre, T. (2016). Fast ventral stream neural activity enables rapid visual categorization. *NeuroImage*, *125*, 280–290. https://doi.org/10.1016/j.neuroimage.2015.10.012

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919. https://doi.org/10.1038/13217

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (in press). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2016.03.063

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*. https://doi.org/10.1038/srep27755

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204. https://doi.org/10.1016/j.neuroimage.2015.07.011

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2016). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology*, jn.00569.2016. https://doi.org/10.1152/jn.00569.2016

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)?(position paper presented to the european cognitive neuropsychology workshop, bressanone, 2005). *Cortex*, *42*(3), 323–331.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … Haxby, J. V. (2012). The Representation of Biological Classes in the Human Brain. *The Journal of Neuroscience*, *32*(8), 2608–2618. https://doi.org/10.1523/JNEUROSCI.5547-11.2012

Contini, E. W., Wardle, S. G., & Carlson, T. A. (in press). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2017.02.013

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/10.1016/S1053-8119(03)00049-1

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4), 16. https://doi.org/10.1167/10.4.16

Crouzet, S. M., & Thorpe, S. J. (2011). Low-Level Cues and Ultra-Fast Face Detection. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00342

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludag, K., De Weerd, P., … Formisano, E. (in press). The impact of ultra-high field MRI on Cognitive and Computational Neuroimaging. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.03.060

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*(5), 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

Desimone, R., Schein, S. J., Moran, J., & Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, *25*(3), 441–452. https://doi.org/10.1016/0042-6989(85)90069-0

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Downing, P. E., Chan, A. W.-Y., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain Specificity in Visual Cortex. *Cerebral Cortex*, *16*(10), 1453–1461. https://doi.org/10.1093/cercor/bhj086

Downing, P. E., & Peelen, M. V. (2016). Body selectivity in occipitotemporal cortex: Causal evidence. *Neuropsychologia*, *83*, 138–148. https://doi.org/10.1016/j.neuropsychologia.2015.05.033

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467.

Edelman, S., & Duvdevani-Bar, S. (1997). Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, *9*(4), 701–720.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321. https://doi.org/10.3758/BF03330618

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, *9*(2), 303–308.

Farah, M. J. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal perception*. Cambridge: MIT Press.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magnetic Resonance Imaging*, *26*(7), 921–934. https://doi.org/10.1016/j.mri.2008.01.052

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4939-2236-9.pdf

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science*, *291*(5502), 312–316. https://doi.org/10.1126/science.291.5502.312

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *The Journal of Neuroscience*, *23*(12), 5235–5246.

Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation Decoding Depends on Maps, Not Columns. *The Journal of Neuroscience*, *31*(13), 4792–4804. https://doi.org/10.1523/JNEUROSCI.5160-10.2011

Freeman, J., Heeger, D. J., & Merriam, E. P. (2013). Coarse-Scale Biases for Spirals and Orientation in Human Visual Cortex. *The Journal of Neuroscience*, *33*(50), 19695–19703. https://doi.org/10.1523/JNEUROSCI.0889-13.2013

Friston, K. J., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *NeuroImage*, *39*(1), 181–205. https://doi.org/10.1016/j.neuroimage.2007.08.013

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220. https://doi.org/10.1002/hbm.460010306

Gaspar, C. M., & Rousselet, G. A. (2009). How do amplitude spectra influence rapid animal detection? *Vision Research*, *49*(24), 3001–3012. https://doi.org/10.1016/j.visres.2009.09.021

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *New York*.

Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychological Science*, *16*(2), 152–160. https://doi.org/10.1111/j.0956-7976.2005.00796.x

Grill-Spector, K., & Malach, R. (2004). The Human Visual Cortex. *Annual Review of Neuroscience*, *27*(1), 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Gross, C. G. (1973). Visual functions of inferotemporal cortex. In R. Jung (Ed.), *Handbook of Sensory Physiology* (pp. 451–482). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-65495-4_11

Gross, C. G. (1994). How Inferior Temporal Cortex Became a Visual Area. *Cerebral Cortex*, *4*(5), 455–469. https://doi.org/10.1093/cercor/4.5.455

Gross, C. G., Rocha-Miranda, C. E. de, & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology*, *35*(1), 96–111.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences*, *111*(10), E962–E971. https://doi.org/10.1073/pnas.1312567111

Hatsopoulos, N. G., & Donoghue, J. P. (2009). The science of neural interface systems. *Annual Review of Neuroscience*, *32*, 249–266.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Haynes, J.-D., & Rees, G. (2005). Predicting the Stream of Consciousness from Activity in Human Visual Cortex. *Current Biology*, *15*(14), 1301–1307. https://doi.org/10.1016/j.cub.2005.06.026

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision*, *8*(12), 9–9. https://doi.org/10.1167/8.12.9

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*(2), 229–289.

Hubel, D. H., & Wiesel, T. N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, *158*(3), 267–293.

Hubel, D. H., Wiesel, T. N., & Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, *177*(3), 361–379.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, *310*(5749), 863–866. https://doi.org/10.1126/science.1117593

Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2016). Typicality sharpens category representations in object-selective cortex. *NeuroImage*, *134*, 170–179. https://doi.org/10.1016/j.neuroimage.2016.04.012

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology*, *115*(4), 2246–2250. https://doi.org/10.1152/jn.01074.2015

Kaiser, D., Oosterhof, N. N., & Peelen, M. V. (2016). The Neural Dynamics of Attentional Selection in Natural Scenes. *Journal of Neuroscience*, *36*(41), 10522–10528. https://doi.org/10.1523/JNEUROSCI.1385-16.2016

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Comput Biol*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309. https://doi.org/10.1152/jn.00024.2007

King, J.-R., & Dehaene, S. (2014a). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130204. https://doi.org/10.1098/rstb.2013.0204

King, J.-R., & Dehaene, S. (2014b). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776. https://doi.org/10.1016/j.visres.2005.10.002

Klein, C. (2010). Philosophical Issues in Neuroimaging. *Philosophy Compass*, *5*(2), 186–198. https://doi.org/10.1111/j.1747-9991.2009.00275.x

Klein, C. (2016). Brain regions as difference-makers. *Philosophical Psychology*, *0*(0), 1–14. https://doi.org/10.1080/09515089.2016.1253053

Konen, C. S., Behrmann, M., Nishimura, M., & Kastner, S. (2011). The Functional Neuroanatomy of Object Agnosia: A Case Study. *Neuron*, *71*(1), 49–60. https://doi.org/10.1016/j.neuron.2011.05.030

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*(1), 577–621.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95.

Mack, M. L., & Palmeri, T. J. (2010). Decoupling object detection and categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1067–1079. https://doi.org/10.1037/a0020254

Mack, M. L., & Palmeri, T. J. (2011). The Timing of Visual Object Categorization. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00165

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Sciences*, *15*(3), 97–103. https://doi.org/10.1016/j.tics.2011.01.004

Mahon, B. Z., Milleville, S. C., Negri, G. A., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron*, *55*(3), 507–520.

Mannion, D. J., McDonald, J. S., & Clifford, C. W. G. (2009). Discrimination of the local orientation structure of spiral Glass patterns early in human visual cortex. *NeuroImage*, *46*(2), 511–515. https://doi.org/10.1016/j.neuroimage.2009.01.052

Marr, D. (1982). *Vision: A computational approach*. MIT Press.

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, *58*(1), 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology*, *100*(3), 1407–1419. https://doi.org/10.1152/jn.90248.2008

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. https://doi.org/10.1016/j.neuroimage.2010.05.051

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. https://doi.org/10.1093/scan/nsn044

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human Object-Similarity Judgments Reflect and Transcend the

Primate-IT Object Representation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00128

Naselaris, T., & Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends in Cognitive Sciences*, *19*(10), 551–554. https://doi.org/10.1016/j.tics.2015.07.005

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 427–436). Retrieved from http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.html

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nowak, L. G., & Bullier, J. (1997). The Timing of Information Transfer in the Visual System. In K. S. Rockland, J. H. Kaas, & A. Peters (Eds.), *Extrastriate Cortex in Primates* (pp. 205–241). Springer US. https://doi.org/10.1007/978-1-4757-9625-4_5

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, 27. https://doi.org/10.3389/fninf.2016.00027

op de Beeck, H. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047

op de Beeck, H., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*(2), 123–135. https://doi.org/10.1038/nrn2314

op de Beeck, H., Torfs, K., & Wagemans, J. (2008). Perceived Shape Similarity among Unfamiliar Objects and the Organization of the Human Object Vision Pathway. *Journal of Neuroscience*, *28*(40), 10111–10123. https://doi.org/10.1523/JNEUROSCI.2511-08.2008

op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, *4*(12), 1244–1252. https://doi.org/10.1038/nn767

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, Statistical, and Practical Perspectives on Pattern-based Classification Approaches to the Analysis of Functional Neuroimaging Data. *Journal of Cognitive Neuroscience*, *19*(11), 1735–1752. https://doi.org/10.1162/jocn.2007.19.11.1735

Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI Activity Dissociate Overlapping Functional Brain Areas that Respond to Biological Motion. *Neuron*, *49*(6), 815–822. https://doi.org/10.1016/j.neuron.2006.02.004

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1, Supplement 1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Philiastides, M. G., & Sajda, P. (2006). Temporal Characterization of the Neural Correlates of Perceptual Decision Making in the Human Brain. *Cerebral Cortex*, *16*(4), 509–518. https://doi.org/10.1093/cercor/bhi130

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Poldrack, R. A. (2010). Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspectives on Psychological Science*, *5*(6), 753–761.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509.

Pratte, M. S., Sy, J. L., Swisher, J. D., & Tong, F. (2016). Radial bias is not necessary for orientation decoding. *NeuroImage*, *127*, 23–33. https://doi.org/10.1016/j.neuroimage.2015.11.066

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/jocn_a_00924

Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2010). Quantifying the Adequacy of Neural Representations for a Cross-Language Phonetic Discrimination Task: Prediction of Individual Differences. *Cerebral Cortex*, *20*(1), 1–12. https://doi.org/10.1093/cercor/bhp076

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., & Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *The Journal of Neuroscience*, *33*(18), 7691–7699. https://doi.org/10.1523/JNEUROSCI.3905-12.2013

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *Neuroimage*, *49*(2), 1545–1558.

Redcay, E., & Carlson, T. A. (2015). Rapid neural discrimination of communicative gestures. *Social Cognitive and Affective Neuroscience*, *10*(4), 545–551. https://doi.org/10.1093/scan/nsu089

Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. *The Journal of Neuroscience*, *34*(26), 8837–8844. https://doi.org/10.1523/JNEUROSCI.5265-13.2014

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–1204. https://doi.org/10.1038/81479

Ritchie, J. B., Bracci, S., & op de Beeck, H. (in press). Avoiding illusory effects in representational similarity analysis: what (not) to do with the diagonal. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2016.12.079

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience*, 190. https://doi.org/10.3389/fnins.2016.00190

Ritchie, J. B., Kaplan, D. M., & Klein, C. (in press). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLoS Comput Biol*, *11*(6), e1004316. https://doi.org/10.1371/journal.pcbi.1004316

Schwarzkopf, D. S., & Rees, G. (2011). Pattern classification using functional magnetic resonance imaging. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 568–579. https://doi.org/10.1002/wcs.141

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *International Conference on Learning Representations*. Retrieved from http://arxiv.org/abs/1312.6229

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Simanova, I., van Gerven, M., Oostenveld, R., & Hagoort, P. (2010). Identifying Object Categories from Event-Related EEG: Toward Decoding of Conceptual Representations. *PLoS ONE*, *5*(12), e14465. https://doi.org/10.1371/journal.pone.0014465

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*. Retrieved from http://arxiv.org/abs/1409.1556

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Spelke, E. S., Phillips, A., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 44–78). New York, NY, US: Clarendon Press/Oxford University Press.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. https://doi.org/10.1007/BF02289729

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *77*, 157–165. https://doi.org/10.1016/j.neuroimage.2013.03.039

Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509. https://doi.org/10.1146/annurev-psych-120710-100412

Torgerson, W. S. (1958). *Theory and methods of scaling* (Vol. xiii). Oxford, England: Wiley.

Ungerleider, L. G. (1982). Two cortical visual systems. *Analysis of Visual Behavior*, 549–586.

van Bergen, R. S., Ji Ma, W., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

Vanrullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Perception Science*, *2*, 365. https://doi.org/10.3389/fpsyg.2011.00365

VanRullen, R. (2017). Perception Science in the Age of Deep Neural Networks. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00142

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, *29*(34), 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage*, *132*, 59–70. https://doi.org/10.1016/j.neuroimage.2016.02.019

Wardle, S. G., & Ritchie, J. B. (2014). Can Object Category-Selectivity in the Ventral Visual Pathway Be Explained by Sensitivity to Low-Level Image Properties? *The Journal of Neuroscience*, *34*(45), 14817–14819. https://doi.org/10.1523/JNEUROSCI.3566-14.2014

Wardle, S. G., Ritchie, J. B., Seymour, K., & Carlson, T. A. (2017). Edge-Related Activity Is Not Necessary to Explain Orientation Decoding in Human Visual Cortex. *Journal of Neuroscience*, *37*(5), 1187–1196. https://doi.org/10.1523/JNEUROSCI.2690-16.2016

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*(6), 685–686. https://doi.org/10.1038/nn1900

Woolgar, A., Golland, P., & Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. *NeuroImage*, *98*, 506–512. https://doi.org/10.1016/j.neuroimage.2014.04.059

Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. *Journal of Cognitive Neuroscience*, *28*(10), 1433–1454. https://doi.org/10.1162/jocn_a_00981

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918. https://doi.org/10.1038/jcbfm.1992.127

Yacoub, E., Harel, N., & Uğurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences*, *105*(30), 10607–10612. https://doi.org/10.1073/pnas.0804110105

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *108*(21), 8850–8855. https://doi.org/10.1073/pnas.1100999108

Chapter 2

Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time-series neuroimaging data

Tijl Grootswagers[1,2,3], Susan G. Wardle[1,2], and Thomas A. Carlson[2,3]

[1] Department of Cognitive Science, Macquarie University, Sydney, Australia

[2] ARC Centre of Excellence in Cognition and its Disorders, Sydney, Australia

[3] School of Psychology, University of Sydney, Australia

Note: This chapter is published as:

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697.

Abstract

Multivariate pattern analysis (MVPA) or brain decoding methods have become standard practice in analysing fMRI data. Although decoding methods have been extensively applied in Brain Computing Interfaces (BCI), these methods have only recently been applied to time-series neuroimaging data such as MEG and EEG to address experimental questions in Cognitive Neuroscience. In a tutorial-style review, we describe a broad set of options to inform future time-series decoding studies from a Cognitive Neuroscience perspective. Using example MEG data, we illustrate the effects that different options in the decoding analysis pipeline can have on experimental results where the aim is to 'decode' different perceptual stimuli or cognitive states over time from dynamic brain activation patterns. We show that decisions made at both preprocessing (e.g., dimensionality reduction, subsampling, trial averaging) and decoding (e.g., classifier selection, cross-validation design) stages of the analysis can significantly affect the results. In addition to standard decoding, we describe extensions to MVPA for time-varying neuroimaging data including representational similarity analysis, temporal generalisation, and the interpretation of classifier weight maps. Finally, we outline important caveats in the design and interpretation of time-series decoding experiments.

1 Introduction

The application of 'brain decoding' methods to the analysis of fMRI data has been highly influential over the past 15 years in the field of Cognitive Neuroscience (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Edelman, Grill-Spector, Kushnir, & Malach, 1998; Haxby et al., 2001; Kamitani & Tong, 2005). In addition to their increased sensitivity, the introduction of fMRI decoding methods offered the possibility to address questions about information processing in the human brain, which have complemented traditional univariate analysis techniques. Although decoding methods for time-series neuroimaging data such as MEG/EEG have been extensively applied in Brain Computing Interfaces (BCI; (Curran & Stokes, 2003; Farwell & Donchin, 1988; Kübler, Kotchoubey, Kaiser, Wolpaw, & Birbaumer, 2001; K.-R. Müller et al., 2008; Vidal, 1973; Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan, 2002), they have only recently been applied in Cognitive Neuroscience (Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011; Duncan et al., 2010; Schaefer, Farquhar, Blokland, Sadakata, & Desain, 2010).

The goal of this article is to provide a tutorial-style guide to the analysis of time-series neuroimaging data for Cognitive Neuroscience experiments. Although introductions to BCI exist (Blankertz, Lemm, Treder, Haufe, & Müller, 2011; Lemm, Blankertz, Dickhaus, & Müller, 2011), the aims of time-series decoding for Cognitive Neuroscience are distinct from those that drive the application of these methods in BCI, thus requiring a targeted introduction. While there are many reviews and tutorials for fMRI decoding (Cox & Savoy, 2003; Formisano, De Martino, & Valente, 2008; Haynes, 2015; Haynes & Rees, 2006; Mur, Bandettini, & Kriegeskorte, 2009; Norman, Polyn,

Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009; Schwarzkopf & Rees, 2011), there are no existing tutorial introductions to decoding time-varying brain activity. Although the approaches are conceptually similar, there are important distinctions that stem from fundamental differences in the nature of the neuroimaging data between fMRI and MEG/EEG. In this paper, we provide a tutorial introduction using an example MEG data set. Although there are many possible analyses targeting time-series data (e.g., oscillatory (Jafarpour, Horner, Fuentemilla, Penny, & Duzel, 2013), or induced responses), we restrict the scope of this article to decoding information from evoked responses, with statistical inference at the group level on single time points or small time windows. As with most neuroimaging analysis techniques, the number of possible permutations for a given set of analysis decisions is very large, and the particular choice of analysis pipeline is guided by the experimental question at hand. Here we aim to provide a broad demonstration of how the analysis may be approached, rather than prescribing a particular analysis pipeline.

Early studies using time-resolved decoding methods have revealed significant potential for experimental investigation using this approach with MEG/EEG (see Section 1.1). However, compared to the popularity of decoding methods in fMRI, to date only a small number of studies have applied multivariate pattern analysis techniques to EEG or MEG. Accordingly, the aims of this article are to (a) Introduce the critical differences between decoding time-series (e.g., MEG/EEG) versus spatial (e.g., fMRI) neuroimaging data; (b) Illustrate the time-series decoding approach using a practical tutorial with example MEG data; (c) Demonstrate the effect that selecting different analysis parameters has on the results, and (d) Outline important caveats in the interpretation of time-series decoding studies. In sum, this article will provide a broad

overview of available methods to inform future time-resolved decoding studies. This tutorial is presented in the context of MEG, however; the methods and analysis principles generalize to other time-varying brain recording techniques (e.g., ECoG, EEG, electrophysiological recordings.). As this review is targeted at providing a broad overview to a general audience, we avoid formal mathematical definitions and implementation details of the methods, and instead focus on the rationale behind the decoding approach as applied to time-series data.

1.1 Multivariate pattern analysis (MVPA) for MEG/EEG

The term 'multivariate pattern analysis' (or MVPA) encompasses a diverse set of methods for analysing neuroimaging data. The common element that unites these approaches is that they take into account the relationships between multiple variables, (e.g., voxels in fMRI, or channels in MEG/EEG), instead of treating them as independent and measuring relative activation strengths. The term ‘decoding’ refers to the prediction of a model from the data (‘encoding’ approaches do the reverse, predicting the data from the model, reviewed in (Naselaris, Kay, Nishimoto, & Gallant, 2011), see also e.g., (Ding & Simon, 2012) for an example of encoding models for MEG). The most common application of decoding in Cognitive Neuroscience is the use of machine learning classifiers (e.g., correlation classifiers (Haxby et al., 2001), or discriminant classifiers (Carlson et al., 2003; Cox & Savoy, 2003)) to identify patterns in neuroimaging data which correspond to the experimental task or stimulus. The most popular applications of MVPA are decoding (for recent reviews on fMRI decoding, see e.g., (Haynes, 2015; Pereira et al., 2009), and more recently, Representational Similarity Analysis (RSA: (Kriegeskorte & Kievit, 2013)). Within the broad category of

MVPA analyses, the central focus of this article is on decoding methods applied to evoked responses, and the increasingly popular RSA framework (Section 5.2).

The decoding approach is illustrated in Figure 1 for a simple experimental design in which the subject viewed pictures of blue circles or red squares while their brain activity was recorded. The goal of the decoding analysis is to test whether we can predict if the subject was viewing a blue circle or red square, based on their patterns of brain activation. If the experimental stimuli can be successfully 'decoded' from the subject's patterns of brain activation, we can conclude that some information relevant to the experimental manipulation exists in the neuroimaging data. First, brain activation patterns in response to the different stimuli (or experimental conditions) are recorded using standard neuroimaging (e.g., MEG, fMRI, etc.) techniques (Figure 1A). The activation levels of the variables (e.g., voxels in fMRI, channels in MEG/EEG) in different experimental conditions are represented as complex patterns in high-dimensional space (each voxel, channel, or principal component is one dimension). For simplicity, in Figure 1B, these patterns are shown in two-dimensional space. Each point in the plot represents an experimental observation corresponding to the simultaneous activation level in two example voxels/channels in response to one of the experimental conditions (blue circles or red squares).

The first step in a decoding analysis involves training a classifier to associate brain activation patterns with the experimental conditions using a subset of the data (Figure 1C). In effect, during training the classifier finds the decision boundary in higher-dimensional space that best separates the patterns of brain activation corresponding to the two experimental categories into two distinct groups. As neuroimaging data is

inherently noisy, this separation is not necessarily perfect (note the red square on the wrong side of the decision boundary in Figure 1C). Next, the trained classifier is used to *predict* the condition labels for new data that was not used for training the classifier (Figure 1D). The classifier predicts whether the new (unlabelled) data is more similar to the pattern of activation evoked by viewing a blue circle or a red square. If the classifier performs higher than that expected by chance (in this case 50% is the guessing rate as there are two stimuli), it provides evidence that the classifier can successfully generalize the learned associations to labelling new brain response patterns. Consequently, it is assumed that the patterns of brain activation contain information that distinguishes between the experimental conditions (i.e., the conditions blue circle/red square can be "decoded" from the neuroimaging data). Decoding accuracy can then be compared across brain regions (in fMRI), or time points (in MEG/EEG), in order to probe the location or time-course of information processing in the brain. This is achieved by repeating the classification multiple times for different data, that is, different time points in MEG/EEG (Figure 1E) for examining the time-course, or different brain regions in fMRI (Figure 1F) for examining the spatial distribution of information in the brain. Thus the main practical differences between decoding from MEG/EEG versus fMRI data lie in the methods used to obtain the patterns of information (Figure 1A, 1B), and the nature of the conclusions drawn from successful decoding performance (Figure 1E, 1F).

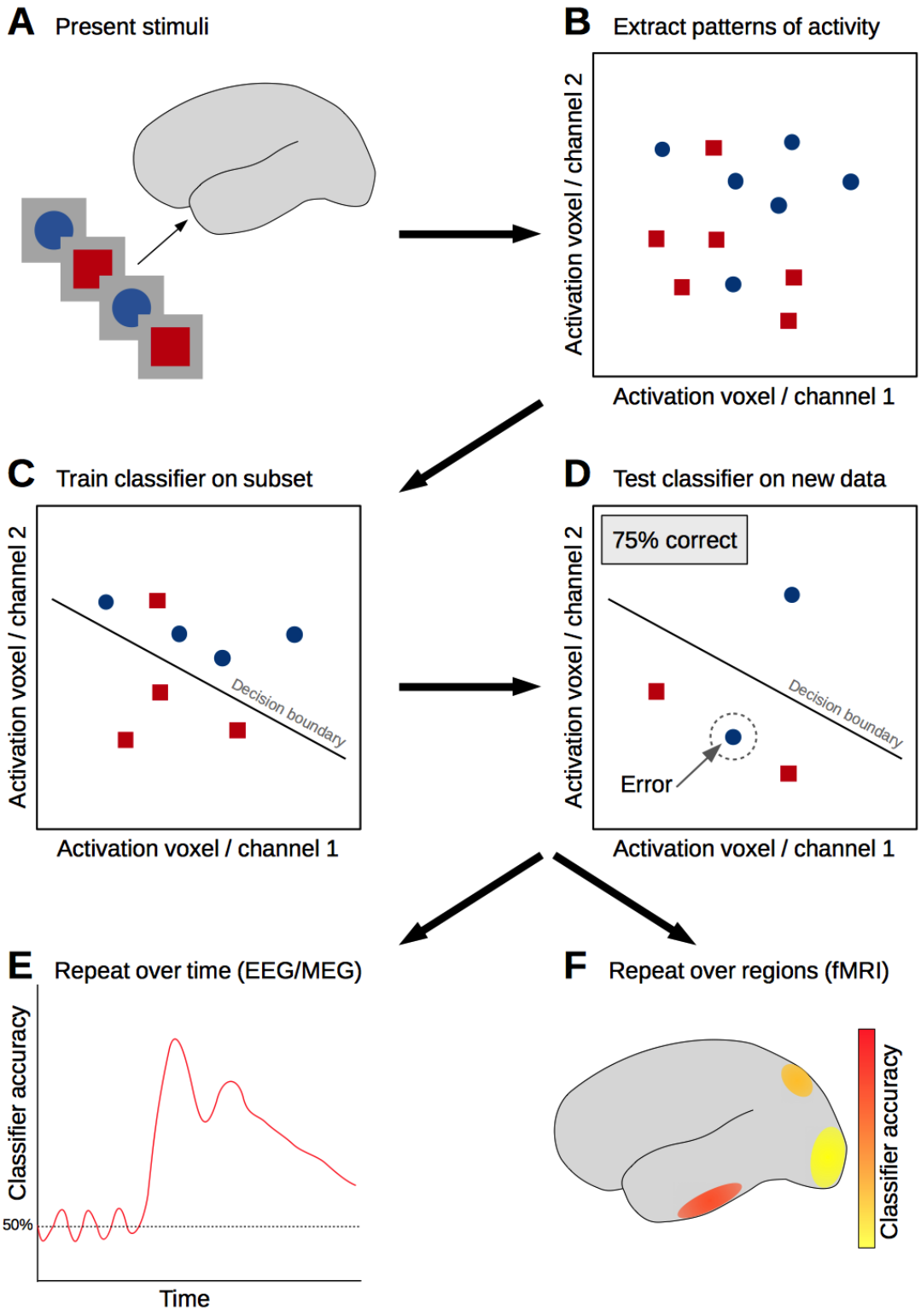
A Present stimuli
B Extract patterns of activity
Activation voxel / channel 2
Activation voxel / channel 1
C Train classifier on subset
Decision boundary
Activation voxel / channel 2
Activation voxel / channel 1
D Test classifier on new data
75% correct
Decision boundary
Error
Activation voxel / channel 2
Activation voxel / channel 1
E Repeat over time (EEG/MEG)
Classifier accuracy
50%
Time
F Repeat over regions (fMRI)
Classifier accuracy

Figure 1. The general decoding approach. A. Brain responses to stimuli (e.g. blue circles and red squares) are recorded with standard neuroimaging techniques. **B.** Patterns of activation evoked by the two stimulus conditions (red square and blue circle) are represented in multiple dimensions (channels in EEG/MEG, or voxels in fMRI); here only two dimensions are illustrated for simplicity. **C.** A classifier is trained on a subset of the neuroimaging data, with the aim of distinguishing a reliable difference in the complex brain activation patterns associated with each stimulus class. **D.** The performance of the classifier in distinguishing between the stimulus classes is evaluated by testing its predictions on independent neuroimaging data (not used in training) to obtain a measure of decoding accuracy. **E,F**. Steps B-D may then be repeated for different time points (when using EEG/MEG) to study the temporal evolution of the decodable signal, or repeated for different brain areas (in fMRI) to examine the spatial location of the decodable information.

Decoding time-series neuroimaging data is becoming increasingly popular. To date, most studies have applied the methods to understanding the temporal dynamics of the processing of visual stimuli and object categories. For example, time resolved decoding has been used to study the emergence of object representations at the category and exemplar level using MEG (Carlson, Tovar, Alink, & Kriegeskorte, 2013), EEG (Cauchoix, Barragan-Jason, Serre, & Barbeau, 2014), and neuronal recordings (Hung, Kreiman, Poggio, & DiCarlo, 2005; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Zhang et al., 2011); how invariant object representations emerge over time (Carlson et al., 2011; Isik, Meyers, Leibo, & Poggio, 2014; Kaiser, Azzalini, & Peelen, 2016); and how objects are represented in other (e.g., written, or auditory) modalities (Chan, Halgren, Marinkovic, & Cash, 2010; Murphy et al., 2011; Simanova, van Gerven, Oostenveld, & Hagoort, 2014, 2010). Other studies have also used this approach to decode the orientation and spatial frequency of gratings from MEG (Cichy, Ramirez, & Pantazis, 2015; Ramkumar, Jas, Pannasch, Hari, & Parkkonen, 2013; Wardle, Kriegeskorte, Grootswagers, Khaligh-Razavi, & Carlson, 2016), and to study

decision making (Bode et al., 2012; Stokes et al., 2013), illusions (Hogendoorn, Verstraten, & Cavanagh, 2015), or working memory (van Gerven et al., 2013; Wolff, Ding, Myers, & Stokes, 2015). Notably, classifiers have been extensively applied to EEG (Guimaraes, Wong, Uy, Grosenick, & Suppes, 2007) for a different goal, as the low cost and portability of EEG is ideal for the development of brain computer interfaces (BCI). These applications use classifiers to predict brain states in order to operate computers or robots (Allison, Wolpaw, & Wolpaw, 2007; Hill et al., 2006; K. Müller, Anderson, & Birch, 2003; K.-R. Müller et al., 2008; Vidal, 1973, p. 2008). However, the goal of BCI is to achieve the maximum possible usability, i.e., optimal prediction accuracy, robust real-time classification, and generalizability. The performance measures of BCI systems are therefore often compared across studies (and in competitions; see e.g., (Tangermann et al., 2012)). This contrasts with decoding in neuroscience, where the goal is to understand brain processing by statistical inference on the availability of information (Hebart, Görgen, & Haynes, 2015), and accuracy differences between studies are generally not taken as meaningful.

Although the field is relatively new, there have already been several methodological extensions to standard decoding analysis applied to time-series neuroimaging data (Section 5). Following its application in fMRI, representational similarity analysis (Kriegeskorte & Kievit, 2013) has been used with MEG data to correlate the temporal structure of brain representations with behaviour (Redcay & Carlson, 2015; Wardle et al., 2016). RSA has also been used to link neuroimaging data from different modalities. For example, for object representations, the representational structure which appears early in the MEG data corresponds to representations in primary visual cortex measured with fMRI, whereas later stages instead reflect the representation in inferior

temporal cortex (Cichy, Pantazis, & Oliva, 2014, 2016). A strength of time-series decoding is that the dynamic evolution of brain representations can be examined. One example of this is the temporal generalization approach (Section 5.1), which has been used in MEG to reveal that local and global responses to auditory novelty exhibit markedly different patterns of temporal generalisation (King, Gramfort, Schurger, Naccache, & Dehaene, 2014). Furthermore, insights into the spatiotemporal dynamics can also be gained by combining source reconstruction methods with the decoding approach (Sudre et al., 2012; van de Nieuwenhuijzen et al., 2013), or by comparing the interaction between subsets of sensors (e.g., (Goddard, Carlson, Dermody, & Woolgar, 2016)). Thus although relatively few time-series neuroimaging studies to date have applied decoding methods, these have already provided valuable insights, illustrating the rich potential for future applications.

Recently, several toolboxes have been developed that implement the methods described in the rest of this paper; The PyMVPA toolbox (Hanke, Halchenko, Sederberg, Hanson, et al., 2009); www.pymvpa.org) handles both fMRI and M/EEG data using the open-source Python language (Hanke, Halchenko, Sederberg, Olivetti, et al., 2009); MNE (Gramfort et al., 2013, 2014); http://martinos.org/mne) is a Python toolbox (and can be accessed in Matlab) designed for M/EEG analyses; the Neural Decoding Toolbox (Meyers, 2013); www.readout.info) is a Matlab toolbox created specifically for time-varying input; and the Matlab toolbox CoSMoMVPA (Oosterhof, Connolly, & Haxby, 2016); www.cosmomvpa.org) handles both fMRI and M/EEG, and was inspired by (and interfaces with) pyMVPA.

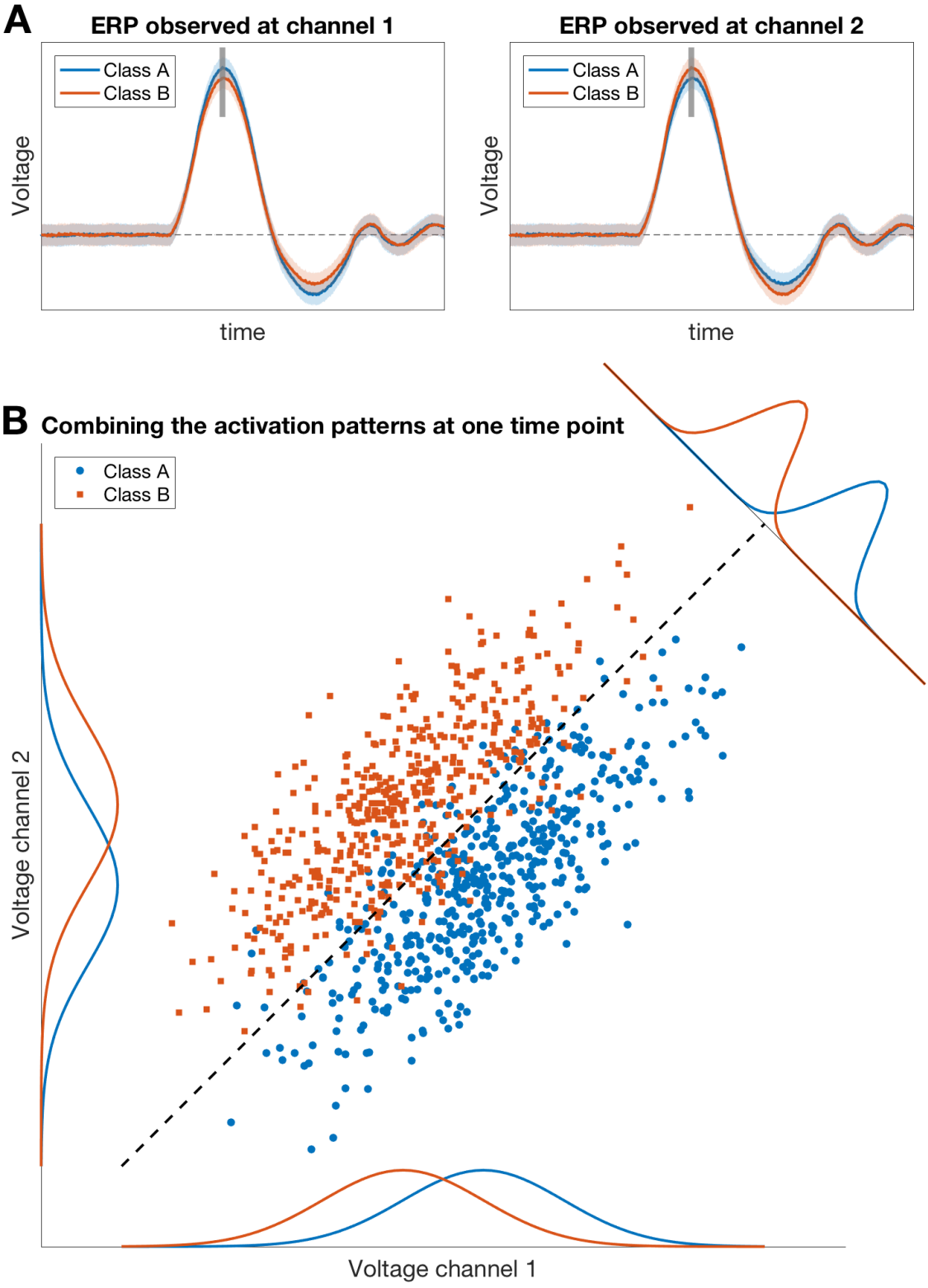
A
ERP observed at channel 1
Class A
Class B
Voltage
time
ERP observed at channel 2
Class A
Class B
Voltage
time
B
Combining the activation patterns at one time point
Class A
Class B
Voltage channel 2
Voltage channel 1

Figure 2. An illustration of how multivariate analysis can result in increased sensitivity compared to univariate analysis. A. Example average event-related potentials (ERPs) in response to two stimuli (class A and class B) are shown in two channels (left and right panels). The responses to the two classes in the individual channels overlap substantially, and potentially non-significant in a univariate analysis. **B.** The same responses represented as points in two-dimensional space, showing the activation in the two channels at one time point (i.e., location of the vertical grey bar in the ERP plots). When combining the information from both channels as in a decoding analysis, it is possible to define a boundary (dashed line) separating the two classes (distributions plotted orthogonal to the dashed line).

Decoding and other variants of MVPA are an alternative and complementary approach to univariate MEG/EEG analysis. This article will not cover univariate methods for MEG and EEG (which are well-established, see e.g., (Cohen, 2014; Luck, 2005)), and as always, the choice of analysis method must be guided by the experimental question. One of the central differences between univariate and multivariate methods is that the classifiers used in decoding approaches can use information that would not be detected when comparing the averaged signals in a univariate analysis (see Figure 2 for an illustration). This can lead to increased sensitivity for detecting differences between conditions (and on a single-trial basis). For example, decoding analysis can result in earlier detection of differences in the signals (Cauchoix, Arslan, Fize, & Serre, 2012; Cauchoix et al., 2014), and the differences found by classifiers can differ from those found in components (Ritchie, Tovar, & Carlson, 2015). Beyond sensitivity, the central distinction between univariate and MVPA analyses are the conceptual differences (activation-based versus information-based) in the experimental questions each approach is suited to addressing. We anticipate that time-series decoding approaches will continue to evolve alongside univariate methods, as has occurred with the adoption of decoding in fMRI, where both methods are used fruitfully.

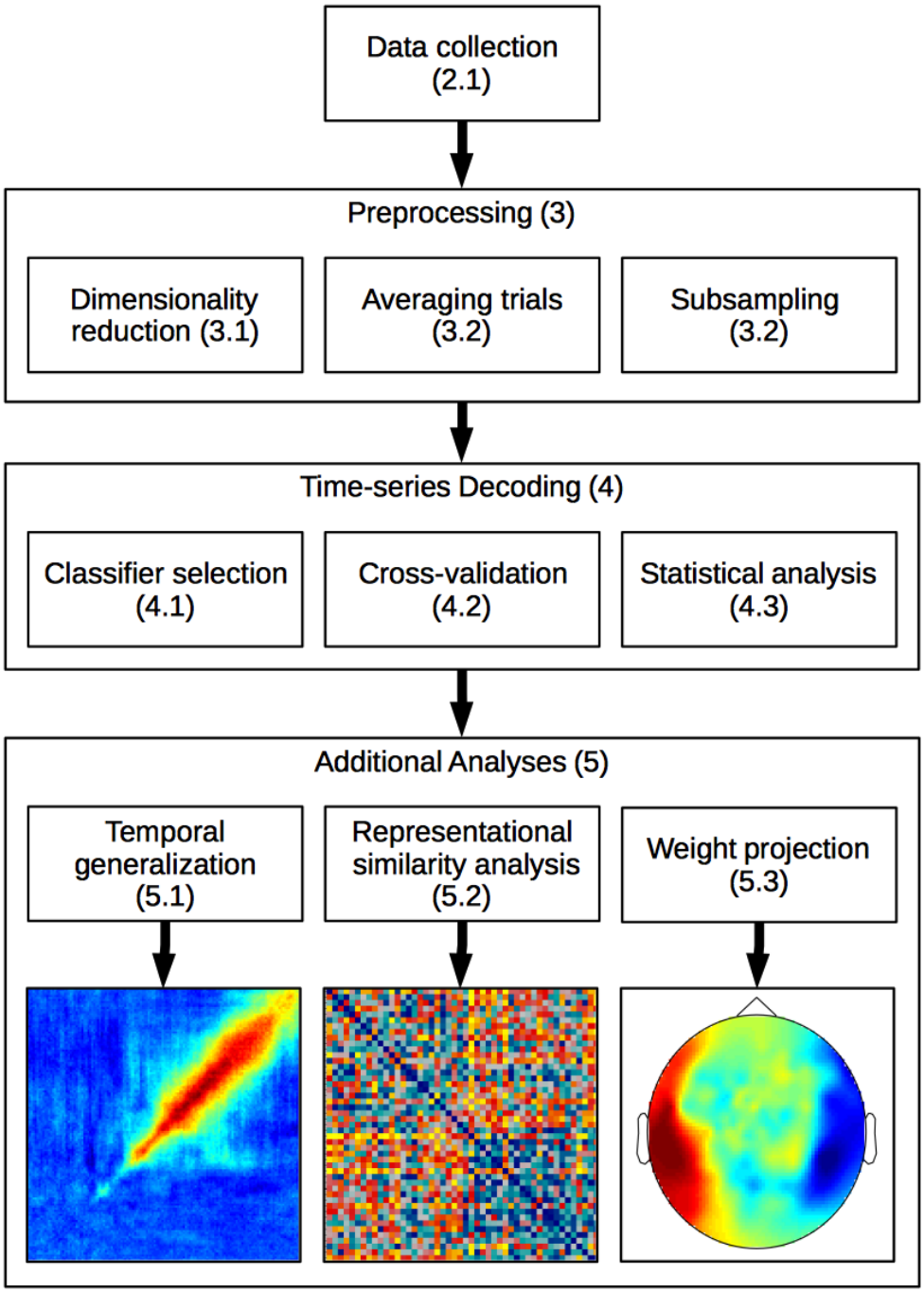

Figure 3. A schematic overview of a typical analysis pipeline. Refer to the relevant sections in the article for further details (numbers in the figure indicate section numbers). This overview illustrates a general pipeline for decoding studies. The practical differences between decoding with MEG/EEG data versus fMRI data arise in both the preprocessing and analysis stages.

The main aim of this article is to describe a typical analysis pipeline for decoding time-series data in a tutorial format. The article is organized as follows; we begin by describing the experiment and the data-recording procedures used to obtain the example MEG data (Section 2). Next, we illustrate how the recordings are preprocessed using a combination of PCA, subsampling and averaging (Section 3). This is followed by the decoding analysis (Section 4). For all analysis stages we provide comparisons of how different choices made at each stage may affect the results. Following the decoding tutorial, in Section 5 we describe three extensions to the method: (1) temporal generalization (King & Dehaene, 2014), (2) representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008), and (3) classifier weights projection (Haufe et al., 2014). Finally, we outline important caveats and limitations of the decoding approach in Section 6. See Figure 3 for an overview of the analysis pipeline and the structure of the paper, including the relevant section numbers.

2 Description of experiment

In this tutorial, we use MEG data to illustrate the effect that different choices made at several analysis stages have on the decoding results. Object animacy has been shown to be a reliably decoded categorical distinction in studies using both fMRI (Downing, Chan, Peelen, Dodds, & Kanwisher, 2006; Kriegeskorte, Mur, Ruff, et al., 2008; Proklova, Kaiser, & Peelen, 2016; Sha et al., 2015) and MEG data (e.g., Carlson et al., 2013; Cichy et al., 2014). Here we use this robust paradigm as a basis for comparing the consequences of different analysis decisions in a decoding pipeline.

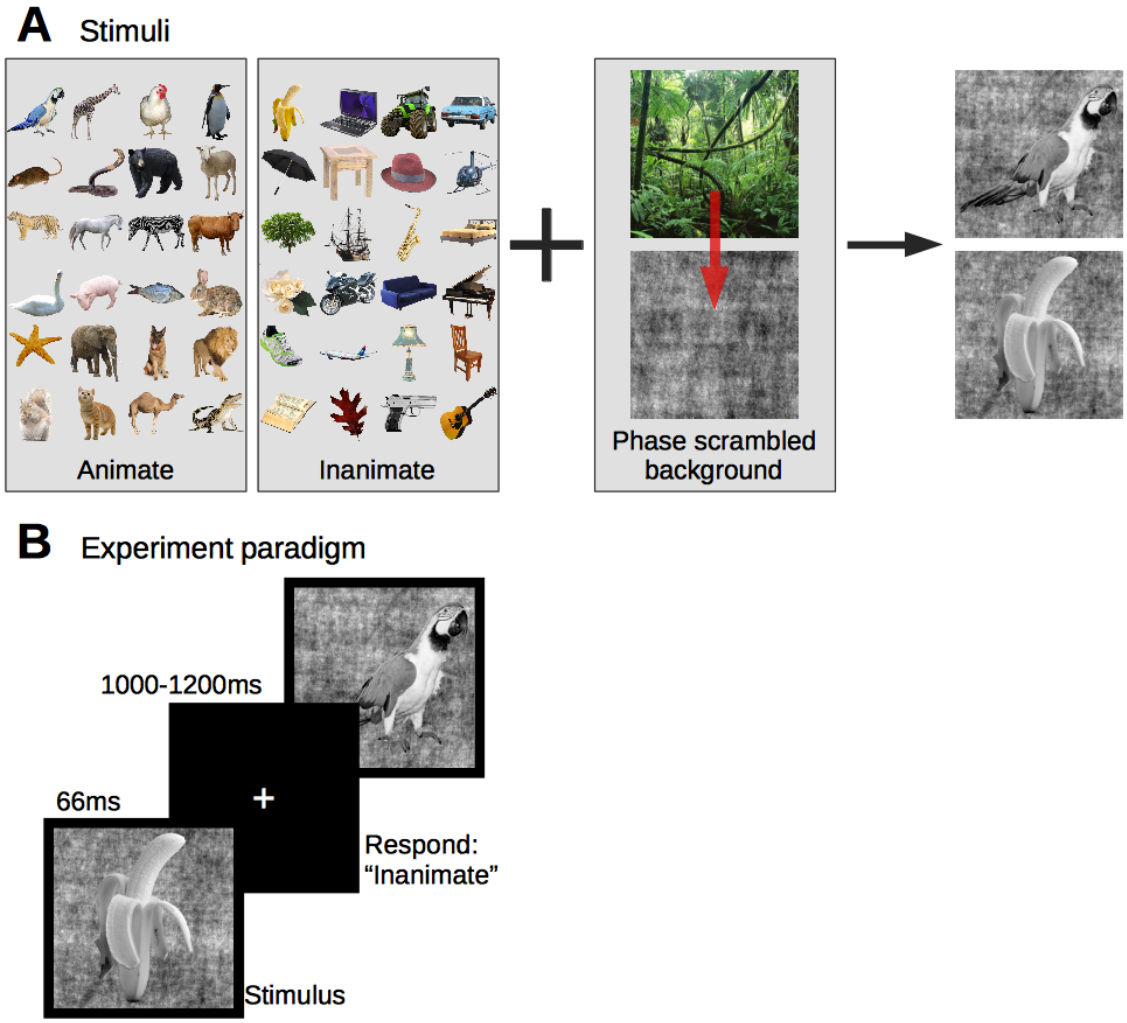

Figure 4. Illustration of the experimental design. A. The stimuli consisted of 24 animate and 24 inanimate visual objects, converted to grey-scale and overlayed on a phase-scrambled natural image background. **B.** Stimuli were presented in random order for 66ms followed by a random ISI between 1000 and 1200ms. Participants categorized the animacy of the stimulus during the ISI with a button press.

Twenty healthy volunteers (4 males) participated in the study with a mean age of 29.3 years (ranging between 24 and 35). Informed consent in writing was obtained from each participant prior to the experiment, and the study was conducted with the approval of the Macquarie University Human Research Ethics Committee. The stimuli were images of 48 visual object exemplars (24 animate and 24 inanimate) segmented and displayed on a phase-scrambled background (See Figure 4) [1]. Stimulus presentation was controlled by custom-written MATLAB (Natick, MA) scripts using functions from Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). The images were shown briefly for 66ms (at 9 degrees visual angle) followed by a fixation cross with a random inter-stimulus interval (ISI) between 1000 and 1200ms. Participants were instructed to categorize the stimulus as 'animate' or 'inanimate' as fast and accurate as possible, using a button press. The response button mapping alternated between 7-minute blocks, to avoid confounding the response with stimulus category (see Section 6.1). This resulted in 32 trials per exemplar, 768 trials per category (animate/inanimate), and 1536 trials total per participant. All trials were included in the analysis, regardless of response, eye blinks or other movement artefacts.

[1] The main study consisted of two conditions, stimuli in a clear or degraded state, however for the purpose of this article we only use the data for stimuli in the clear state (normal photographs of objects).

2.1 Data Collection

The MEG signal was continuously sampled at 1000Hz from 160 axial gradiometers[2] using a whole-head MEG system (Model PQ1160R-N2, KIT, Kanazawa, Japan) inside a magnetically shielded room (Fujihara Co. Ltd., Tokyo, Japan) while participants lay in a supine position. Recordings were filtered online with a high-pass filter of 0.03Hz and a low-pass filter of 200Hz. The recordings were imported into MATLAB using the Yokogawa MEG Reader Toolbox for MATLAB (YOKOGAWA Electric Corporation, 2011). The first step in the pipeline was to slice the data into epochs (i.e., trials), time-locked to a specific event. We extracted -100ms to 600ms of MEG data relative to the stimulus onset. The first 100ms of signal taken prior to trial onset serves as a sanity check for decoding accuracy (see Section 4.2).

2.2 Analysis Summary

The effect of different choices on the decoding results will be described by systematically varying one parameter relative to a set of fixed parameters. Three caveats of this approach are that (1) as these parameters are not independent, interactions between analysis decisions are likely, (2), the effects of these analysis decisions will vary between data sets, and (3), drawing conclusions on differences in decoding performances is only valid when the noise level is the same in all cases. Consequently, the following results should be interpreted as illustrative, rather than provide prescriptive analysis guidelines. All analysis code for the examples was written

[2] Other MEG systems also include magnetometers, and there are possible differences in decodability from gradiometers and magnetometers (Kaiser, Azzalini, & Peelen, 2016).

in Matlab (Natick, MA), using only standard functions unless otherwise specified. In order to illustrate the effects of different parameters on the results, they are consistently shown at the final stage plotted as a function of classifier accuracy over time. The default methods and fixed parameters are listed here for reference, and unless otherwise specified, the results in Figures 6-10 are obtained using this default pipeline:

- Preprocessing (Section 3)
    - Subsampling 200Hz
    - Averaging 4 trials
    - PCA retaining 99% of the variance
- Decoding (Section 4)
    - Naïve Bayes classifier
    - Leave-one-exemplar-out cross-validation

The results are reported as time-varying decoding accuracy, i.e., higher accuracies reflect better decoding (prediction) of stimulus animacy from the MEG data. To assess whether accuracy was higher than chance, a Wilcoxon signed-rank test on the grand mean of decoding performance (N=20) was performed at each time point. The resulting p-values were corrected for multiple comparisons by controlling the false discovery rate (FDR, (Benjamini & Hochberg, 1995)). Note that these statistics were chosen for their simplicity and ease of use, we discuss commonly used options for assessing classifier performance and statistics in Section 4.3.

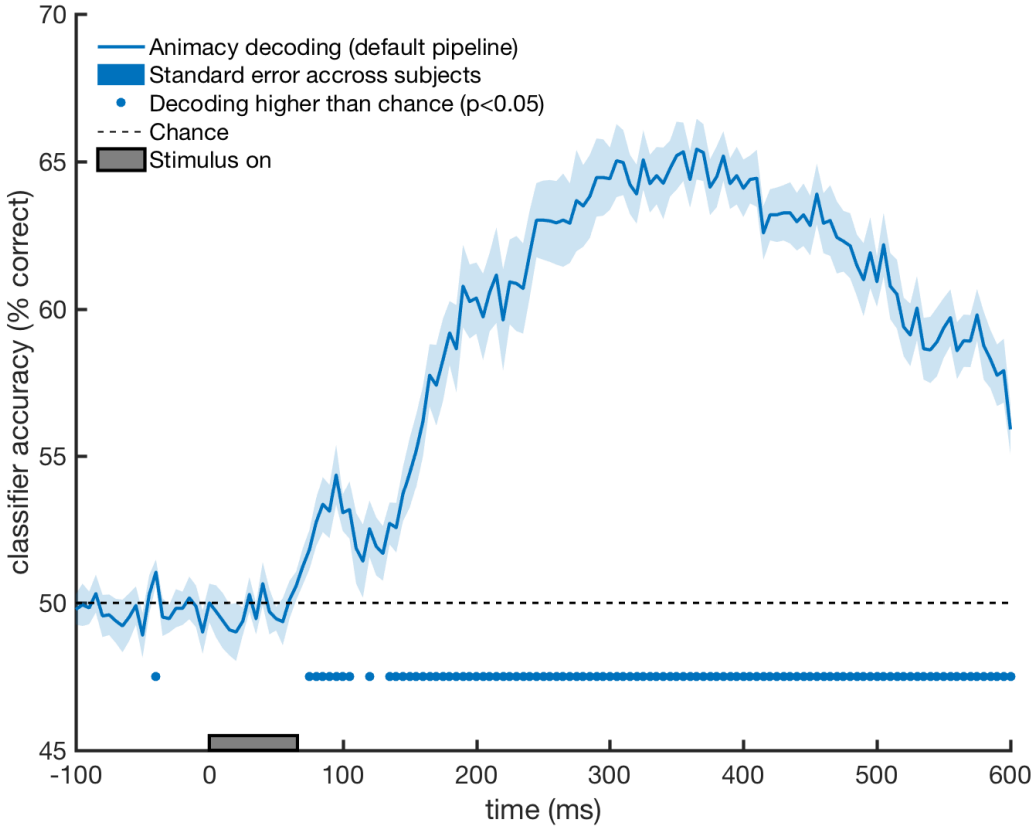

Figure 5. Decoding animacy from MEG data using the default analysis pipeline. Classifier accuracy (percent correct averaged across subjects) is shown as a function of time relative to stimulus onset at 0ms. The dashed line marks chance classification accuracy at 50%. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

Figure 5 shows the result of this default pipeline. As expected, before stimulus onset (-100 to 0ms), decoding performance is at chance (50%), confirming that there is no animacy information present in the signal. Then, approximately 80ms after stimulus presentation, the classifier's performance rises significantly above chance for almost the entire time window (to 600ms). Thus, at these time points, we are able to successfully decode from the MEG activation patterns whether the presented stimulus in a given trial was animate (e.g., parrot, dog, horse, etc.), or inanimate (e.g., banana, chair, tree, etc.). This indicates that the MEG signal contains information related to the animacy of the stimulus. The next sections will describe this pipeline in detail while comparing the effect of different analysis decisions.

3 Preprocessing

Neuroimaging data is often noisy. The signals in imaging data are weak compared to, for example, environmental noise, baseline activity levels, or fluctuations caused by eye blinks or other movements. Therefore, a set of standard procedures is used to increase the signal-to-noise ratio. Furthermore, neuroimaging data are high-dimensional, and it is common practice to restrict the analysis to fewer dimensions. In MEG decoding, the dimensions of the data are generally reduced in the number of features (i.e., channels) that are input to the classifier. In addition, temporal smoothing is commonly applied. There are multiple ways to achieve these preprocessing steps, the most common are described in this section.

3.1 Data transformation and dimensionality reduction

A standard step in preprocessing is to reduce the dimensionality of the data. Some classifiers require more training samples than features; and others might overfit to noise in the data if provided with too many features (Bishop, 2006; De Martino et al., 2008; Misaki, Kim, Bandettini, & Kriegeskorte, 2010), or require longer computation time. Raw MEG recordings consist of many channels, typically 160 or more, and there is considerable redundant information, e.g., in adjacent channels. It is therefore common practice to reduce the dimensionality of the data by feature selection prior to decoding, which can be accomplished in multiple ways. One approach is to select the channels that are most informative (De Martino et al., 2008; Hanke, Halchenko, Sederberg, Hanson, et al., 2009). (Isik et al., 2014) for example, by using an ANOVA significance test to select the MEG channels that contain significant stimulus-specific information.

Alternatively, one can use unsupervised, data-driven approaches such as Principal Component Analysis (PCA), which transforms the data into linearly uncorrelated components with the same number of feature dimensions, ordered by the amount of variance explained by each component (For a detailed introduction to PCA, see (Jackson, 1991)). The use of PCA for MEG has a number of advantages: First, retaining only the components that account for most of the variance substantially reduces the dimensionality of the data. In the example data (160 channels), on average 48.16 (*SD*=7.05, range: 26 - 79) components accounted for 99% of the variance in the data. Secondly, PCA can separate out noise and artefacts such as eye blinks (see section 3.2) into their own components. These components can then be

suppressed by the classifier because they do not contain class-specific information. Third, as the resulting PCA components are uncorrelated, it allows for using simpler (i.e., faster) classifiers that assume no feature covariance (e.g., Naïve Bayes, see Section 4.1).

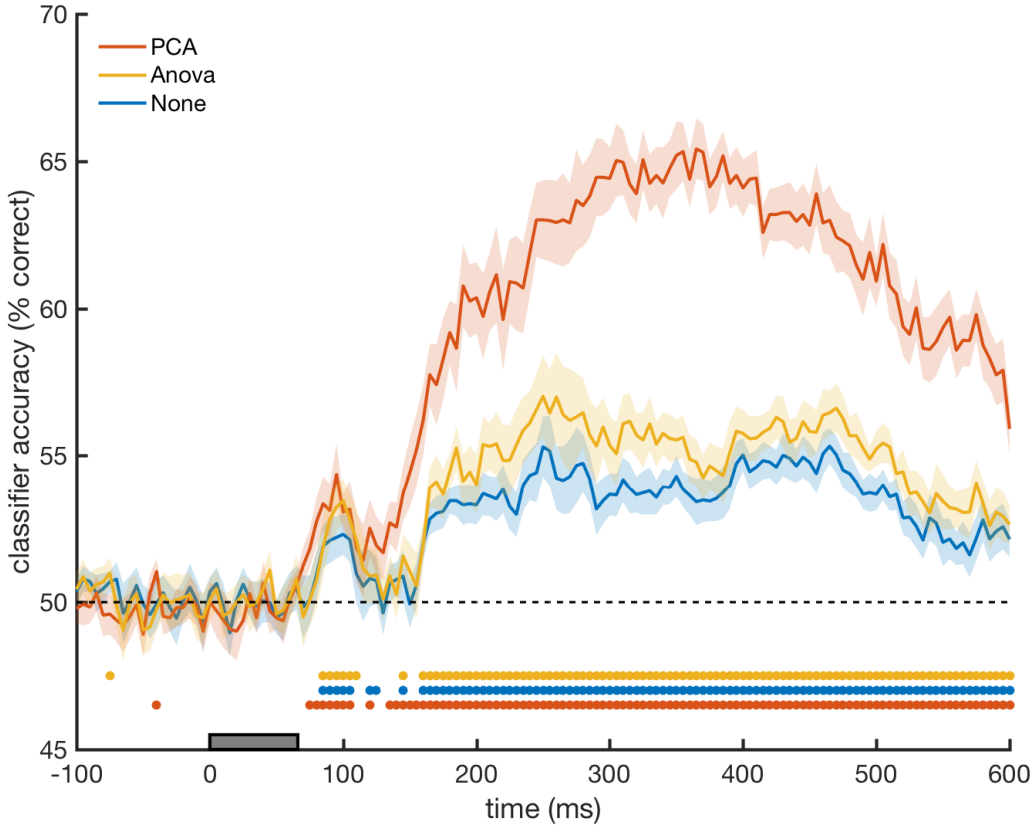

Figure 6. The effect of dimensionality reduction methods on decoding performance. The effect of channel selection using ANOVA (yellow line) is marginally better than using the raw data (blue line). Using PCA (red line) yields the largest gain in performance. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

Figure 6 illustrates the effect of the described dimensionality reduction methods on decoding performance for the example data. For this data set and classifier, PCA yields much better performance compared to using the raw channels (Isik et al., 2014). Note that these differences are classifier dependent (as shown in Section 4.1). Here, the PCA transformation was computed on the training data, and applied on the test data, separately for each time point, and separately for each training fold. Alternatively, one could compute *one* transformation for the whole time-series, and/or do this on *all* data before the cross-validation process. However, this is only viable if the goal of the analysis is statistical inference (Hebart et al., 2015), as it could result in more optimistic decoding accuracies that would not generalize to new data[3].

An alternative method is to transform the sensor-level data into activations in virtual source space. Instead of decoding channel-level activations, source reconstruction (e.g., beamformer (Van Veen, Van Drongelen, Yuchtman, & Suzuki, 1997), or minimum norm estimate (Hämäläinen & Ilmoniemi, 1994)) can be applied during preprocessing. Classification is then performed in source space rather than channel space (Sandberg et al., 2013; Sudre et al., 2012; van de Nieuwenhuijzen et al., 2013). Using source space for decoding has the potential to improve classification accuracies (Sandberg et al., 2013; van de Nieuwenhuijzen et al., 2013), as source reconstruction algorithms can ignore channel-level noise. Inferences about the spatial origin of the decoded discrimination can be made by restricting the classifier to considering signals from pre-determined regions of interest (Sudre et al., 2012), or by using the complete

[3] Note that when comparing PCA performed inside the cross-validation loop on separate time points with PCA performed before the cross-validation on all time points, we did not find any difference in classifier accuracy (data not shown), but this may not hold for different data sets.

source space reconstruction and projecting the classifier weights (see Section 5.3) into source space (van de Nieuwenhuijzen et al., 2013). The second approach relies on interpreting classifier weights, and therefore, the reliability of the sources depends not only on the reconstruction quality, but also on decoding performance (see Section 5.3). Source reconstruction methods are still developing, and reconstruction accuracies are likely to improve in the future, making source space decoding an attractive option. However, as source space decoding has not been widely used to date, we will not cover it in the rest of this tutorial.

3.2 Improving signal to noise

MEG data is generally sampled at high frequencies (e.g., 1000Hz), and a common strategy to improve signal-to-noise (the strength of the signal compared to the strength of the background noise) is by collapsing data over time. The two main approaches are to classify on more than one time point using a sliding window (e.g., Ramkumar et al., 2013), or down-sample the data to lower frequencies (see Figure 7). The difference between the methods is that when using a sliding window, the classifier has access to all time points in the window (the number of features is increased), while in subsampling, it receives the average (the number of features at each time point stays the same). For the example data, subsampling has a small effect on decoding performance, but also benefits the analysis by reducing the computation time for the decoding analysis as there are fewer time points to classify. The sliding window approach also improves performance, but the benefit is marginal especially considering that the computation time increases significantly with larger sliding windows as the classifier is still trained and tested at each time point. The optimal parameters will

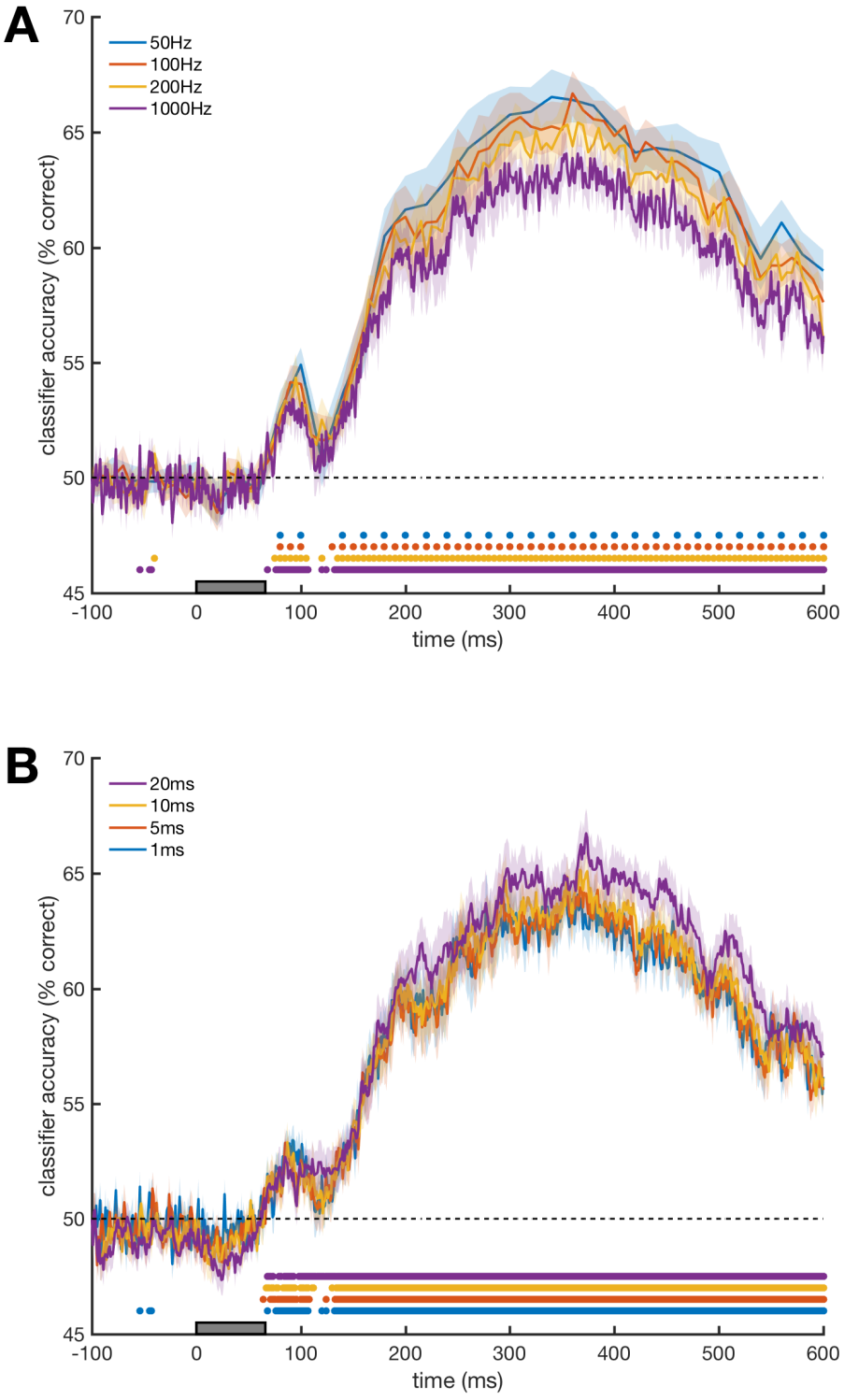

Figure 7. The effect of (A) subsampling and (B) sliding window approaches to improving signal-to-noise on classifier accuracy. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

depend on the particular data set and desired temporal resolution. An important caveat for both approaches is that estimates of both decoding onset and the time of peak decoding are affected by the choice of subsampling or sliding window. When using a sliding time window, the last time bin in the window should be used for determining the onset (as in Figure 7), to avoid shifting the onset forward in time. It is recommended to apply a low-pass filter before resampling (e.g., subsampling using the *decimate* function in MATLAB) as subsampling can cause aliasing. Low-pass filtering, however, can cause an artefact whereby significant decoding emerges even when no signal exists in the original data (Vanrullen, 2011). For the example data, we subsampled by a factor of 5 to obtain a sampling rate of 200Hz.

Another source of noise originates from artefacts. Eye blinks, eye movements, heartbeats, and muscle movement can cause significant artefacts. Typically, in classical M/EEG analyses trials containing such artefacts are manually inspected and excluded from the analysis, or independent component analysis is used to separate out these artefacts into their own components, which are then removed manually or automatically (Mognon, Jovicich, Bruzzone, & Buiatti, 2011). Experiments can also be designed in a way to reduce the number of artefacts, for example by instructing participants to blink in response to a particular stimulus that is not part of the analysis (Cichy et al., 2014). We did not perform any artefact rejection on our data, and found classification performance to be well above chance, but this can vary across data sets. As classifiers have the capacity to learn to ignore bad channels or supress noise during training, artefact correction is likely less critical in decoding analyses. However, note that if artefacts are confounded with a condition (e.g., if more eye movements occurred in one condition than the other due to some property of the stimulus), this would make

the artefacts a potential source of discrimination information for the classifier. If this is the case, it would not be possible to determine whether the classifier was decoding the experimental condition, or the correlated difference in artefacts (see also Section 6.1).

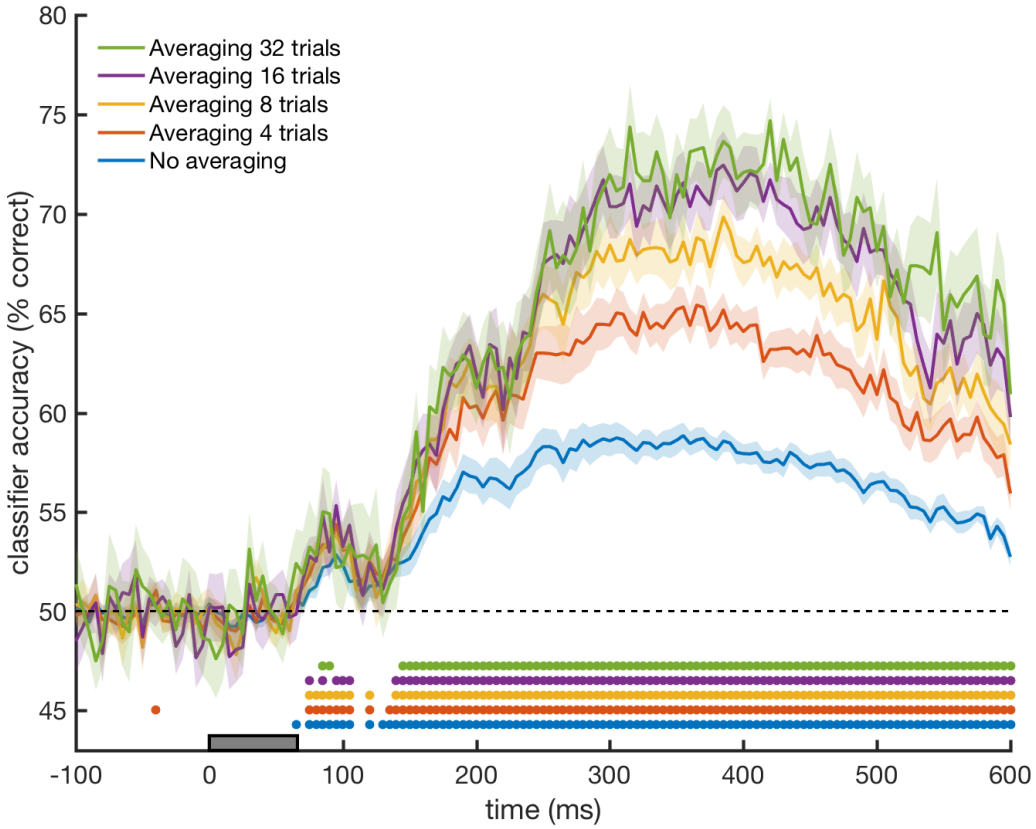

Figure 8. The effect of averaging trials on decoding performance. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

Increased signal-to-noise can also be achieved by averaging trials belonging to the same exemplar before decoding (Isik et al., 2014). Averaging increases general decoding performance and makes signatures (e.g., onsets, maxima or minima) more pronounced. This effect is shown in Figure 8, where different numbers of trials (belonging to the same exemplar) are averaged. Interestingly, the first onset of

decoding is similar regardless of the number of trials that are averaged. The greatest increase in performance (in our example data) is observed when averaging 4 trials. Averaging more trials does not increase decoding performance by the same factor, suggesting that here 4 trials is a good trade-off between signal-to-noise, and trials per exemplar. The trade-off to consider when selecting the number of trials to average is that reducing the trials per exemplar (e.g., averaging 32 trials here produces only one trial per exemplar) typically increases the variance in (within-subject) classifier performance. Alternatively, when not enough trials are available, the trials used for training the classifier could be sampled with replacement (bootstrapped). The optimal number of trials to average will differ for different data (e.g., in (Isik et al., 2014), averaging 10 trials was used). Note that trial averaging does not affect model testing (e.g., RSA, Section 5.2), as relative decoding performance is scaled similarly between exemplars or time points.

4 Decoding

Decoding analysis is performed on the preprocessed data. To summarize, in preprocessing the raw MEG signal is sliced into epochs from -100 to 600ms relative to stimulus onset, then down-sampled to 200Hz. Groups of 4 single trials are averaged to boost signal-to-noise, resulting in 8 pseudo-trials for each object exemplar. These preprocessed pseudo-trials are the input to the classifier in the decoding analysis.

In order to decode the class information (animacy) from the MEG data, a pattern classifier (see Section 4.1) is trained to distinguish between two classes of stimuli (animate and inanimate objects). The classifier's ability to generalize this distinction to

new data is assessed using cross-validation (see Section 4.2). If the classifier's performance after cross-validation is significantly above chance, this indicates that the MEG patterns contain class-specific information, and we conclude that the class can be decoded from the MEG data. In time-resolved MEG decoding studies, this process is repeated on all time points in the data. Then, for example, one can examine when the peak in decoding performance occurs, i.e., at what time point the information in the signal allows for the best class distinction. Another feature often used is the onset of significant decoding performance, to determine the earliest time that class-specific information becomes available. These signatures can then be compared across experimental conditions.

4.1 Classifiers

There are numerous types of classifiers, which originate from the machine learning literature. Classifier choice has the potential to influence experimental results, as different classifiers make different assumptions about the data. In addition, the goal of classification in machine learning is high predication accuracy, which drives the development of increasingly sophisticated classifier algorithms. In contrast, prediction is not the main goal of decoding in neuroscience, and classifier choice instead favours simplicity and ease of interpretation over optimizing prediction accuracies. Therefore, for brain decoding studies, linear classifiers are generally preferred, as they are simpler in nature, making interpretation less complex (Misaki et al., 2010; K. Müller et al., 2003; Schwarzkopf & Rees, 2011). The default classifiers used in fMRI decoding are typically linear support-vector machines (SVM), or, to a lesser extent, correlation classifiers. However, fMRI data typically has many features/dimensions. SVM is generally better

than other classifiers when dealing with many features and is therefore a popular choice. In comparison to fMRI data, time-series data often has fewer features (e.g., our example MEG data set uses only ~50 components following PCA). Consequently, it is possible that there are differences in the suitability of different classifiers for fMRI versus time-series decoding analysis. Here we compare the performance of SVM, correlation classifiers, and two common alternatives (Linear Discriminant Analysis (LDA) and Gaussian Naïve Bayes (GNB)) on the example MEG data (Figure 9), using their built-in Matlab implementations (and default parameters). Notably, LDA, GNB and SVM have the best overall performance. Taking the complexity of the classifier into account, which affects the computational requirements and given that classification is generally repeated many times (e.g., on multiple time points), this argues in favour of the discriminant classifiers (GNB and LDA), which are faster to train than SVM. Interestingly, despite their relative popularity in fMRI, the correlation classifiers did not perform as well on our data. However, (Isik et al., 2014) reported correlation classifier performance for their MEG data on par with other classifiers. This difference could be due to many factors, for example, different choices in the preprocessing pipeline or experimental design. To illustrate that classifier performance depends on preprocessing, we tested the same classifiers using different preprocessing decisions. For example, Figure 9B shows that not performing PCA has a large effect on GNB performance, but a smaller effect on the performance of LDA and SVM. These dependencies highlight the difficulty in attempting to make universal recommendations for decoding analyses. Furthermore, each classifier has a number of parameters that may be optimised, however, most neuroscience studies use standard classifier implementations.

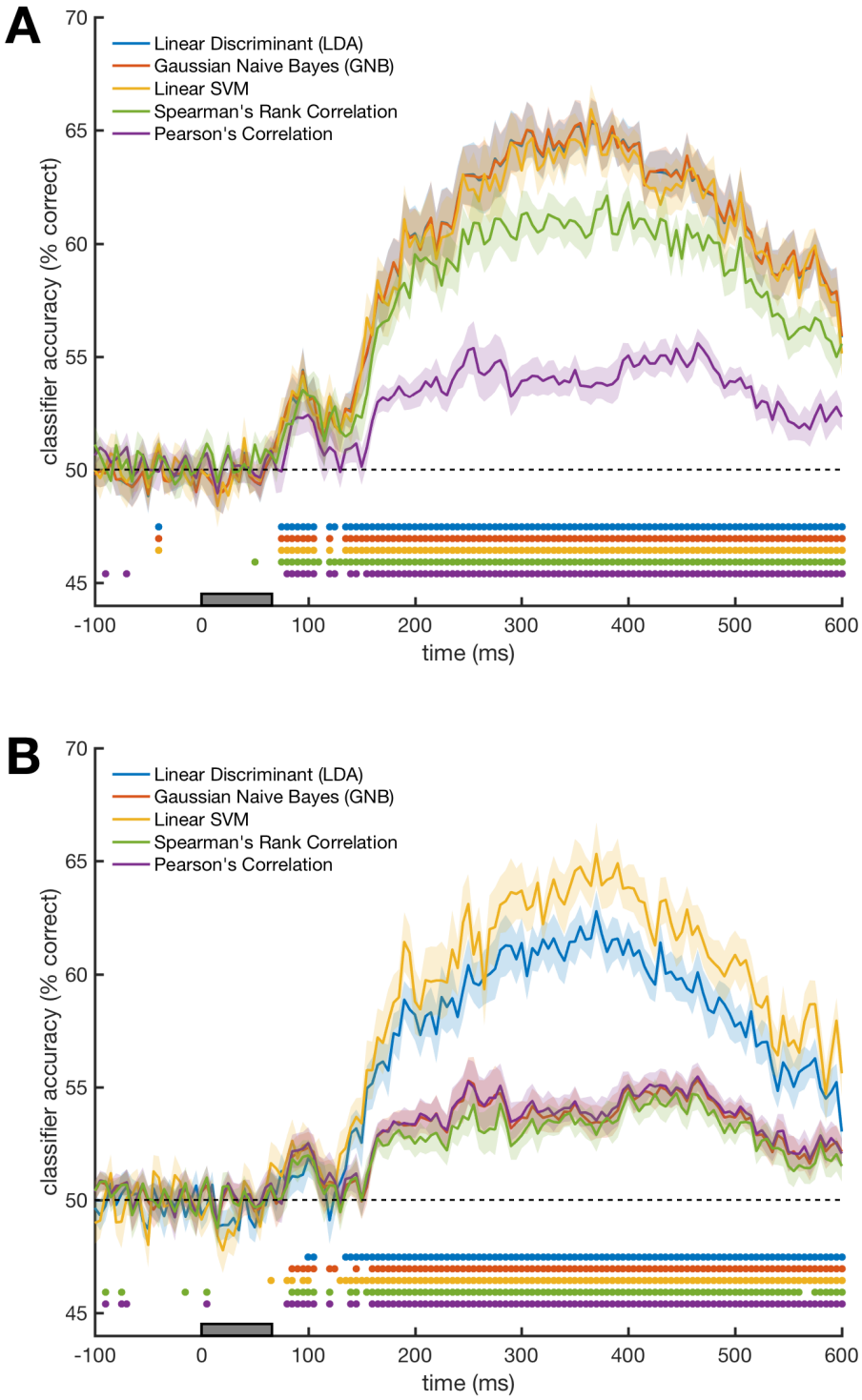

Figure 9. Comparison of classification accuracy as a function of classifier type. **A.** Using the standard decoding pipeline. **B.** Using the standard pipeline without performing PCA. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

4.2 Cross-validation

An essential step in decoding analysis is cross-validation: this provides an evaluation of classifier generalization performance. In standard *k*-fold cross-validation, the data is divided into *k* subsets (i.e., folds), where each subset contains a balanced amount of trials from each class (e.g., animate and inanimate exemplars in our example experiment). The classifier is trained using all-but-one subsets (the training set). Next, the trained classifier is used to predict the class of the trials from the remaining subset (the test set). This process is repeated for all subsets, and the average classifier performance across all folds is reported. This method makes maximal use of the available data, as all trials are used for testing the classifier. Note that in fMRI decoding the sets are often based on experimental runs (leave-one-*run*-out cross validation), as the trials within each run are not independent (e.g., due to the slow hemodynamic response). In MEG decoding, individual trials are generally assumed to be independent (Oosterhof et al., 2016), and trials are randomly assigned to train and test sets. The theoretical optimal performance is obtained by leave-one-*trial*-out cross-validation, where the classifier is trained on all-but-one trial. It is however computationally more intensive, especially with many trials (which is typically the case in MEG).

As with other analysis decisions, the most appropriate implementation of cross-validation is guided by the experimental design. Standard *k*-fold cross-validation assigns individual trials to training and testing sets. Depending on the research question, this may produce a confound in the class distinction that the classifier learns from the training data. For example, for decoding animacy, standard cross-validation would entail that trials belonging to the same exemplar (e.g., ‘car’) are assigned to both

training and test sets. Consequently, it may be possible for the classifier to learn to distinguish the classes based on the activation patterns evoked by visual properties of specific exemplars. This makes it unclear whether the classification boundary is based on animacy or visual features. To avoid this, when decoding categories composed of many exemplars; we recommend leave-one-*exemplar*-out cross-validation (see Carlson et al., 2013), where all trials belonging to one exemplar (e.g., car) are assigned to the test set and the classifier is trained on the data from the other exemplars (e.g., 'dog' and 'chair'). This is repeated for all exemplars (i.e., every exemplar is assigned to the test set once).

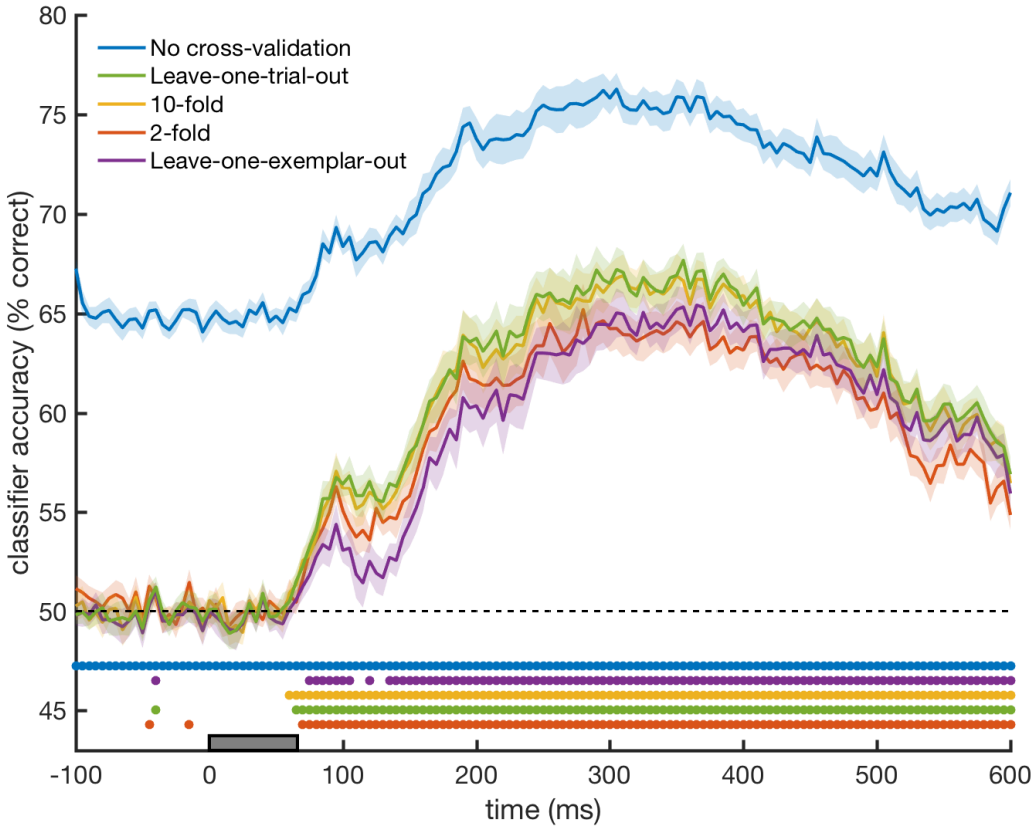

Figure 10. Classification accuracy as a function of cross-validation method. The shaded area is the standard error across subjects. Discs above the x-axis indicate the time points where decoding performance is significantly higher than chance.

Figure 10 shows decoding accuracy for different forms of cross validation, including an invalid analysis without cross validation. Note that without cross validation, classifier performance is above chance *prior* to stimulus onset. This nonsensical result arises from the test data being used to train the classifier, violating the constraint of independence. Time-resolved decoding methods have a convenient built-in check for this: above chance decoding performance before stimulus onset suggests an error exists in either the preprocessing or cross validation stages. In our data, 10-fold and leave-one-trial-out cross validation yielded very similar results, suggesting that the optimal split is data-specific. Further, by comparing performance between traditional cross validation (e.g., *k*-fold) and leave-one-*exemplar*-out, it is possible to estimate to what degree classifier performance is driven by individual stimulus properties (e.g., low-level visual properties of the exemplar images). The difference between *k*-fold and leave-one-exemplar-out cross validation is observed early in the time-series (consistent with the timing of early visual feature processing), and is reduced later in the time course (Figure 10). Taken together, a valid form of cross-validation with independent training and test data is essential. Although there are several ways of splitting up the data into training and test sets, the particular version of cross-validation implemented must be compatible with the research question.

4.3 Evaluation of classifier performance and group-level statistical testing

Statistical evaluation of decoding analyses is a complex issue, and there is not yet consensus on the optimal approach (Allefeld, Görgen, & Haynes, 2016; Nichols & Holmes, 2002; Noirhomme et al., 2014; Schreiber & Krekelberg, 2013; Stelzer, Chen, & Turner, 2013). The statistical approach used in our example analysis is common in

the literature (e.g., (Carlson, Tovar, et al., 2013; Ritchie et al., 2015)) and was chosen for its simplicity; however, there are several alternative methods that are also valid. For example, we report classifier performance as accuracy (percent correct). Accuracy is a less appropriate measure when dealing with unbalanced data (more trials exist for one class than for the other), as a trained classifier could exploit the uneven distribution and achieve high accuracy simply by predicting the more frequent class. For unbalanced data, a measure of performance that is unaffected by class bias such as D-prime is more appropriate. Alternatively, 'balanced accuracy' includes the mean of the accuracies for each class and thus is also unaffected by any class imbalance in the data.

Several options exist for assessing whether classifier performance is significantly above chance. The non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945) was used in our example (Carlson, Tovar, et al., 2013; Ritchie et al., 2015)), as it makes minimal assumptions about the distribution of the data. Alternatively, the Student's t-test is also commonly used (but see (Allefeld et al., 2016). Another popular alternative is the permutation test, which entails repeatedly shuffling the data and recomputing classifier performance on the shuffled data to obtain a null-distribution, which is then compared against observed classifier performance on the original set to assess statistical significance (see e.g., (Cichy et al., 2014; Isik et al., 2014; Kaiser et al., 2016). Permutation tests are especially useful when no assumptions about the null-distribution can be made (e.g., in the case of biased classifiers or unbalanced data), but they take much longer to run (e.g., repeating the analysis ~10,000 times).

Importantly, as is the case in fMRI analyses, time-series neuroimaging analyses also require addressing the problem of multiple comparisons (Bennett, Baird, Miller, & Wolford, 2011; Bennett, Wolford, & Miller, 2009; Nichols, 2012; Pantazis, Nichols, Baillet, & Leahy, 2005) as typically multiple tests are conducted across different time points. The FDR adjustment used in our example analysis is straightforward, but a limitation is that it does not incorporate the relation between time points (Chumbley & Friston, 2009). Alternatively, cluster-based multiple-comparison correction involves testing whether *clusters* of time points show above-chance decoding and therefore can result in increased sensitivity to smaller, but more sustained effects (Mensen & Khatami, 2013; Nichols, 2012; Oosterhof et al., 2016; Smith & Nichols, 2009).

5 Additional analyses

In the sections above, we illustrated the standard approach to decoding time-series neuroimaging data. Here we outline three extensions for decoding analysis. The first is temporal cross-decoding (Section 5.1), which tests the degree to which activation patterns in response to the experimental conditions are sustained or evolve over time. The second is the RSA framework (Section 5.2), which facilitates the testing of models of the structure of decodable information over time. Finally, we outline a method that involves projection of the classifier weights in order to determine the spatial source of the signal driving the classifier in sensor-space (Section 5.3).

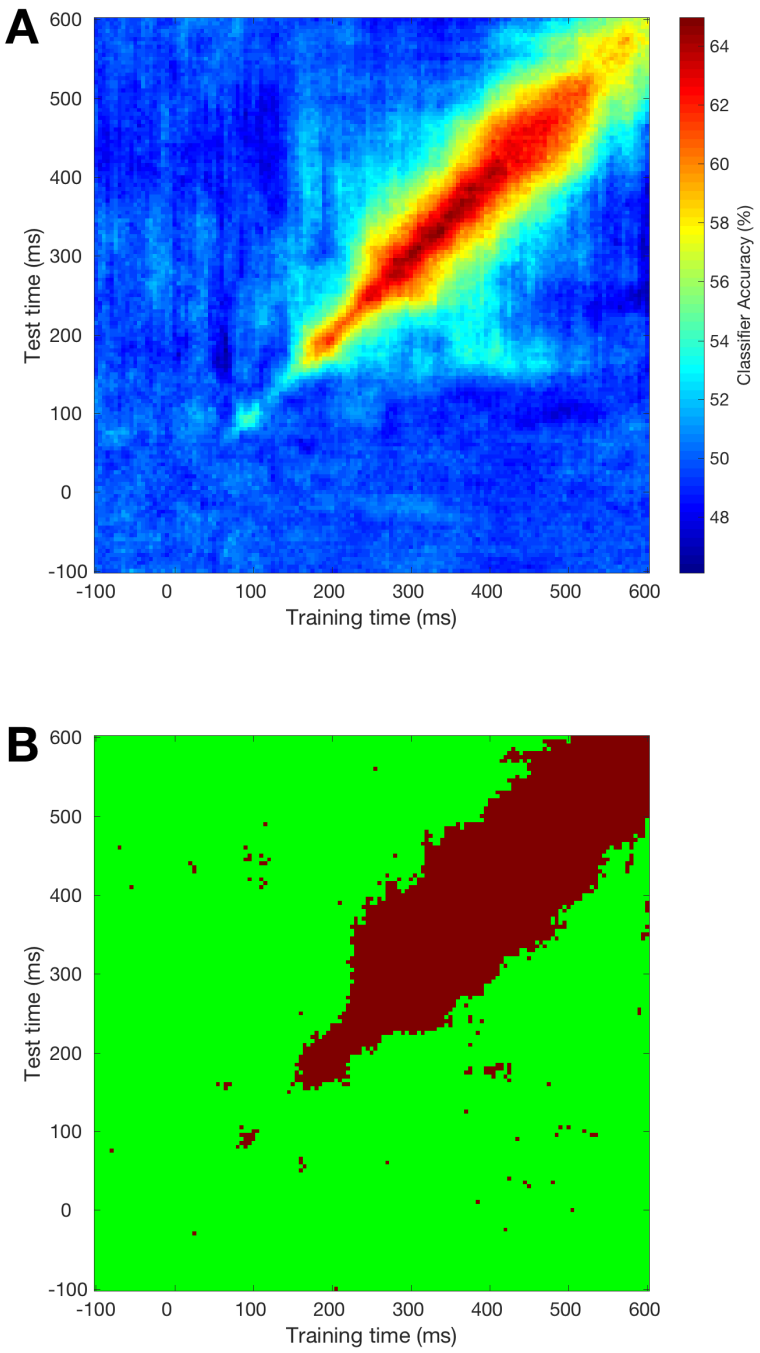

Figure 11. Temporal generalization A. Temporal generalization of decoding performance. A classifier is trained at one time point, and tested at a different time point. This is repeated for all pairs of time points. The figure shows the generalization accuracy averaged over subjects. **B.** Map of time point pairs where the generalization was significantly different (red area) from chance (Wilcoxon signed rank test, controlled for multiple comparisons using FDR).

5.1 The temporal generalization method

An advantage of time-series decoding is that it has the potential to reveal the temporal evolution of brain activation patterns, rather than providing a single, static estimate of decodability for a stimulus or task. One method is to train a classifier on a particular time point, and then test its decoding performance on *different* time points. This form of cross-decoding reveals to what degree the activation patterns for a particular stimulus or task evolve. Classifiers effectively carve up multidimensional space in order to distinguish between the experimental conditions, thus when a classifier which is trained on one time point can successfully predict class-labels for data at other time points, it suggests that the structure of the multidimensional space is similar across time. Conversely, if cross-decoding is unsuccessful across two time points, it suggests that the multidimensional space has changed sufficiently for the boundary between classes determined at one time point to be no longer meaningful by the second time point. Beyond temporal characterisation of the decoding results, this method has the potential utility to test cognitive models which make theoretical predictions about the generalizability of representations (see also Figure 4 in (King & Dehaene, 2014). For example, the temporal generalization of classifiers can be tested between two completely separate datasets. (Isik et al., 2014) tested the temporal generalization performance of a classifier that was trained on stimuli that were presented foveally, and then tested on peripherally presented stimuli. Similarly, (Kaiser et al., 2016) used this method to distinguish category-specific responses from shape-specific responses.

Figure 11A shows cross-validated temporal cross-decoding performed on the example MEG data. The diagonal in this figure is analogous to the standard one-dimensional

time-series decoding plot (e.g., Figures 5-10). Significant points (shown in Figure 11B) off the diagonal indicate that the classifier, when trained on data from time point A, can generalize to data from time point B. The generalization accuracy normally drops off systematically away from the diagonal. In this case, classifier performance generalizes well for neighbouring time points (red region on the diagonal) as expected, and additionally, to some extent between 150-200 and 300-500ms, indicating that the MEG activation patterns are similar in these windows.

5.2 Representational Similarity Analysis (RSA)

Standard decoding analysis reveals whether class-specific information is present in the neuroimaging signal. Approaches such as cross-decoding (e.g., temporal generalisation) can begin to probe the underlying representational structure of the information in the brain activation patterns used by the classifier. RSA takes this concept further, and provides a framework for testing hypotheses about the structure of this information (Kriegeskorte, Mur, & Bandettini, 2008). RSA is based on the assumption that stimuli with more similar neural representations are more difficult to decode. Conversely, stimuli with more distinct representations are expected to be easier to decode. Thus the central idea is that representational similarity can be indexed by the degree of decodability. By comparing the decodability of all possible pair-wise combinations of stimuli, a representational dissimilarity matrix (RDM) is calculated. That is, for each pair of stimuli, the distance between their activation patterns is computed using one of several distance metrics (e.g., correlation between the activation patterns, or difference in classifier performance (Walther et al., 2016).

An example RDM is shown in Figure 12A, in which each cell in the matrix corresponds to the dissimilarity of two of the object stimuli in the MEG animacy experiment. For data with high temporal resolution such as MEG, a series of RDMs can be created for each time point, and used to investigate the temporal dynamics of representations over time. The time-varying RDMs in Figure 12A are constructed by decoding all pairwise stimuli using the same pipeline (using 2-fold cross-validation, as leave-one-exemplar out is not possible when decoding between 2 exemplars), thus one square in the RDM represents the decoding accuracy for classifying between one pair. Following calculation of the RDM (either time-varying or static) from the empirical data, the empirical RDM can be compared to model RDMs that make specific predictions about the relative decodability of the stimulus pairs. In RSA studies to date, model RDMs have been constructed from predictions based on a wide range of sources: including behavioural results, computational models, stimulus properties, or neuroimaging data from a complementary imaging method such as fMRI (e.g., (Carlson, Simmons, Kriegeskorte, & Slevc, 2013; Cichy et al., 2014, 2016; Kriegeskorte, Mur, Ruff, et al., 2008; Redcay & Carlson, 2015; Wardle et al., 2016).

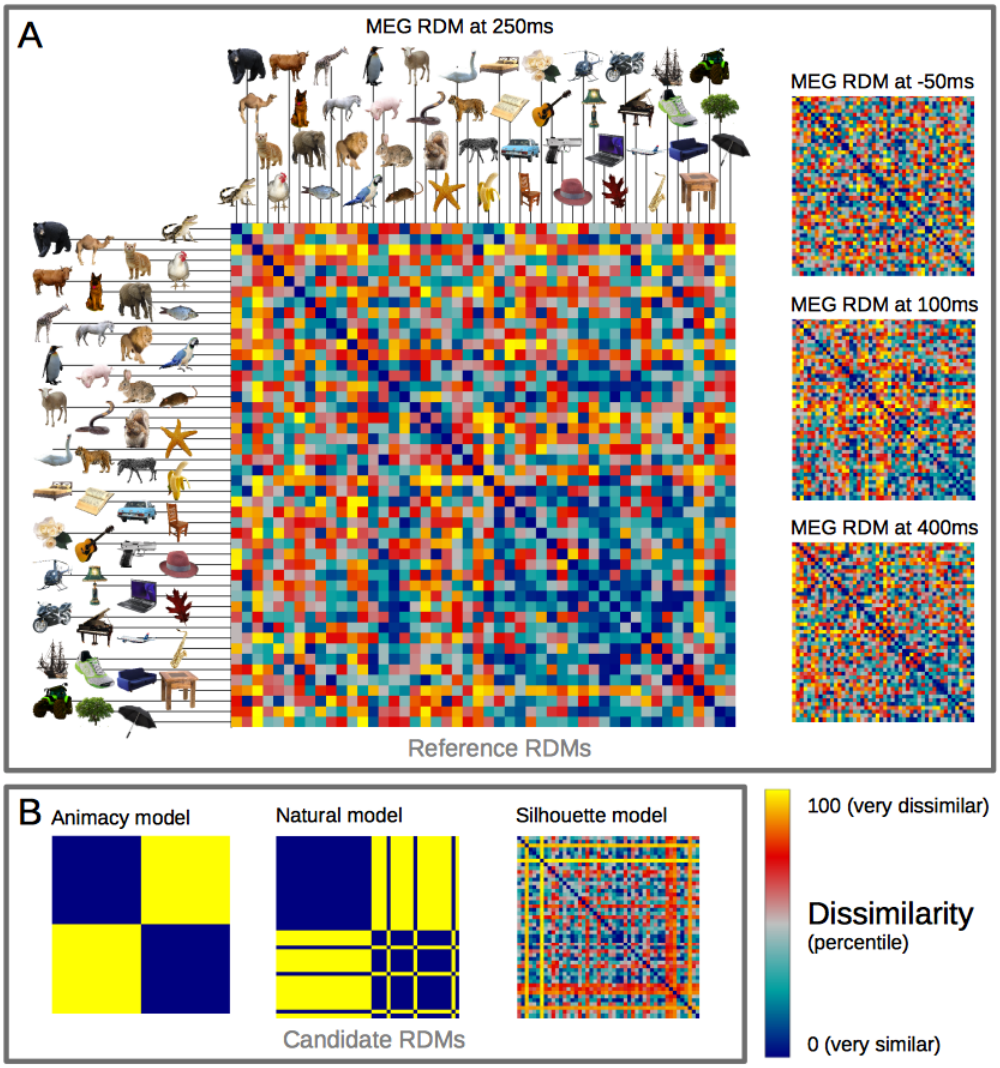
A
MEG RDM at 250ms
MEG RDM at -50ms
MEG RDM at 100ms
MEG RDM at 400ms
Reference RDMs
B
Animacy model
Natural model
Silhouette model
Candidate RDMs
100 (very dissimilar)
Dissimilarity
(percentile)
0 (very similar)
C
Animacy model
Natural model
Silhouette model
Noise ceiling (lower bound)
Mean RDM correlation (Kendall's tau)
0.1
0.08
0.06
0.04
0.02
0
-0.02
-100
0
100
200
300
400
500
600
time (ms)

Figure 12. Model evaluation within the RSA framework. A. The empirical MEG RDMs averaged across subjects. One cell in the matrix represents the dissimilarity between the MEG activation patterns for one pair of object exemplars. RDMs are shown for four time points: -50ms, 100ms, 250ms, and 400ms. **B.** Three model RDMs, which predict the representational similarity of the brain activation patterns for all object pairs based on different stimulus properties: an Animacy model (Animate vs. Inanimate objects), a Natural model (Natural vs. Artificial objects), and a Silhouette model (based on the visual similarity of the objects' silhouettes). **C.** RSA model evaluation. At each time point, the empirical RDMs for each subject are correlated with the three candidate model RDMs in B. The strength of the average correlations shows how well the candidate models fit the data. Shaded areas represent the standard error over subjects, and the marks above the x-axis indicate time points where the mean correlation was significantly higher than zero (Wilcoxon signed-rank test, controlled for multiple comparisons using FDR). The grey dotted line represents the lower bound of the 'noise ceiling' at each time point, which is the theoretical lower bound of the maximum correlation of any model with the reference RDMs at each time point, given the noise in the data (Nili et al., 2014).

Figure 12 shows the results of RSA model evaluation for the example MEG data. For each time point, the empirical RDMs (Figure 12A) are correlated with three theoretical models (Figure 12B); a model of stimulus animacy, a model that distinguishes artificial versus natural stimuli, and a control model based on the visual similarity of the exemplar's silhouettes (which correlates well with early stimulus discriminability, see e.g., (Carlson et al., 2011; Redcay & Carlson, 2015). Each of these models predicts the relative (dis)similarity of the MEG activation patterns for each exemplar pair based on their specific stimulus features. The extent of the correlation between the model and empirical MEG RDMs is interpreted as reflecting the degree to which the 'representational structure' characterised by each model exists in the brain activation patterns. The results in Figure 12C are plotted as the correlation between the three model RDMs with the MEG RDM over time. The Animacy model (blue line) has a

better fit to the MEG data than the Natural model (orange line), and both models have a better fit than the Silhouette model (yellow line) later in the time series. The Silhouette model has the best fit early in the time series, which is expected as it represents early visual features. This suggests that animacy is a relatively good predictor of the similarity of the MEG activation patterns for the exemplar pairs: object pairs from the same category (e.g., both animate) are more difficult to decode than object pairs from different categories (e.g., one animate and one inanimate). Within the RSA framework, this is interpreted as evidence that animacy is a key organising principle in the representational structure of the object exemplars.

Despite its strengths, a current limitation of the RSA approach is that valid statistical comparison of different candidate models is difficult (Kriegeskorte & Kievit, 2013; Thirion, Pedregosa, Eickenberg, & Varoquaux, 2015). A recent development proposes evaluating model performance by comparing it to the highest possible performance given the noise in the data, called the 'noise ceiling' (Nili et al., 2014). When applied to MEG data, the performance of various models relative to the noise ceiling (computed from the empirical data as described in (Nili et al., 2014) can be evaluated over time, as shown in Figure 12C. Despite the present limitations in directly comparing different models, RSA is a useful tool for investigating the structure of the decodable signal in neuroimaging data, which will undoubtedly continue to evolve in its sophistication and utility. For a more detailed introduction, see (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014)).

5.3 Weight projection

Following successful classification of experimental conditions, it is sometimes of interest to examine the extent to which different voxels (fMRI) or sensors (MEG/EEG) drive classifier performance. During standard classification analysis, each feature (e.g., MEG sensors) is assigned a weight corresponding to the degree to which its output is used by the classifier to maximize class separation. Therefore, it is tempting to use the raw weight as an index of the degree to which sensors contained class-specific information. However, this is not straightforward, as higher raw weights do not directly imply more class-specific information than lower weights. Similarly, a non-zero weight does not imply that there is class-specific information in a sensor (for a full explanation, proof, and example scenarios,(Haufe et al., 2014). This is because sensors may be assigned a non-zero weight not only because they contain class-specific information, but also when their output is useful to the classifier in suppressing noise or distractor signals (e.g., eyeblinks or heartbeats). An elegant solution to this issue was recently introduced by (Haufe et al., 2014) and has been applied to MEG decoding (Wardle et al., 2016). This consists of transforming the classifier weights back into activation patterns. Following this transformation, the reconstructed patterns are interpretable (i.e., non-zero values imply class-specific information) and can be projected onto the sensors. It is import to note however, that the reliability of the patterns depends on the quality of the weights. That is, if decoding performance is low, weights are likely sub-optimal, and reconstructed activation patterns have to be interpreted with caution (Haufe et al., 2014).

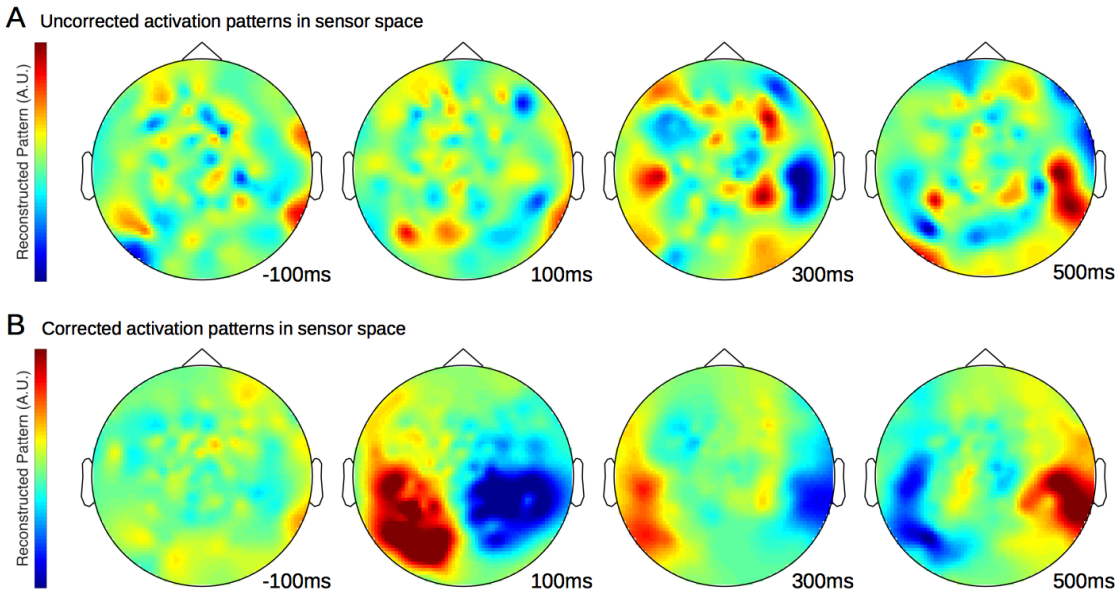

Figure 13. Classifier weights projected onto MEG sensor space. The corresponding time points are shown beneath the scalp topographies. Darker colours indicate channels that contribute to animacy decoding. **A.** Uncorrected (raw) weights projections cannot be interpreted directly, as classifiers can assign non-zero weights to channels that contain no class-specific information. **B.** The activation patterns computed from transformed weights (following the method of (Haufe et al., 2014)) can be interpreted.

Here we summarize this transformation for MEG data, and plot the results in Figure 13. First, the classifier weights (we used LDA instead of GNB in this example as this method only applies to classifiers that consider the feature covariance) are transformed into activation patterns by multiplying them with the covariance in the data: $A = \mathrm{cov}(X)*w$; where X is the NxM matrix of MEG data with N trials and M features (channels), and w is a classifier weight vector of length M. A is the resulting vector of length M containing the reconstructed activation patterns (i.e., the transformed classifier weights). For display purposes, the reconstructed activation patterns can be projected onto the scalp location of the channels. Figure 13B shows the result for the example MEG data at four time points (using the FieldTrip toolbox for MATLAB:

(Oostenveld, Fries, Maris, & Schoffelen, 2010); here the results are scaled by the inverse of the source covariance ($A*cov(X*w)^{-1}$) to allow for comparison across time points. Note that this method cannot be directly used if multiple time points are used for classification (e.g., the sliding window approach described in Section 3.2). The uncorrected (raw) weight projections are shown for comparison in Figure 13A. We can now observe that for the activation patterns in Figure 13B, the information source is located approximately around the occipital lobes (back sensors) at 100ms, and later around the temporal lobes (side sensors) at 300ms, as expected from the visual processing hierarchy. Notably, this pattern is not as easily identifiable in the raw weight topographies shown in Figure 13A. For an in-depth explanation (with examples) of the weights interpretation problem and its solution, see (Haufe et al., 2014).

6 General discussion

Time-series decoding methods provide a valuable tool for investigating the temporal dynamics and organization of information processing in the human brain. In the previous sections we outlined an example decoding analysis pipeline for time-series neuroimaging data, illustrated effects of different methods and parameters (and their interactions), and introduced extensions of the method such as temporal generalisation (5.1), RSA (5.2), and weights projection (5.3). In the final section, we discuss some important aspects to consider when performing these analyses and interpreting the results. One of the central issues concerns the interpretation of classifier accuracy. Classifiers are extremely sensitive and will exploit all possible information in the data. This means that careful experimental design and interpretation of the results is required in order to draw meaningful conclusions from decoding studies (see e.g.,

(Carlson & Wardle, 2015; de-Wit, Alexander, Ekroll, & Wagemans, 2016; Naselaris & Kay, 2015). The next section outlines a number of such pitfalls to avoid in the implementation of time-series decoding methods.

6.1 Common pitfalls

The first caveat applies to all studies using classifiers and is well-described in the literature (Kriegeskorte, Lindquist, Nichols, Poldrack, & Vul, 2010; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Pereira et al., 2009). It is important that the classifier has no access to class-specific information about the data contained in the test set, as this will artificially inflate classifier performance. This analysis confound is referred to as 'double dipping', and was demonstrated in the analysis without cross-validation in Figure 10 (Section 4.2). One advantage of time-series decoding is that in most cases, data obtained before stimulus onset serves as a first check. If classifier accuracy is above chance before stimulus onset, it indicates possible contamination from double dipping.

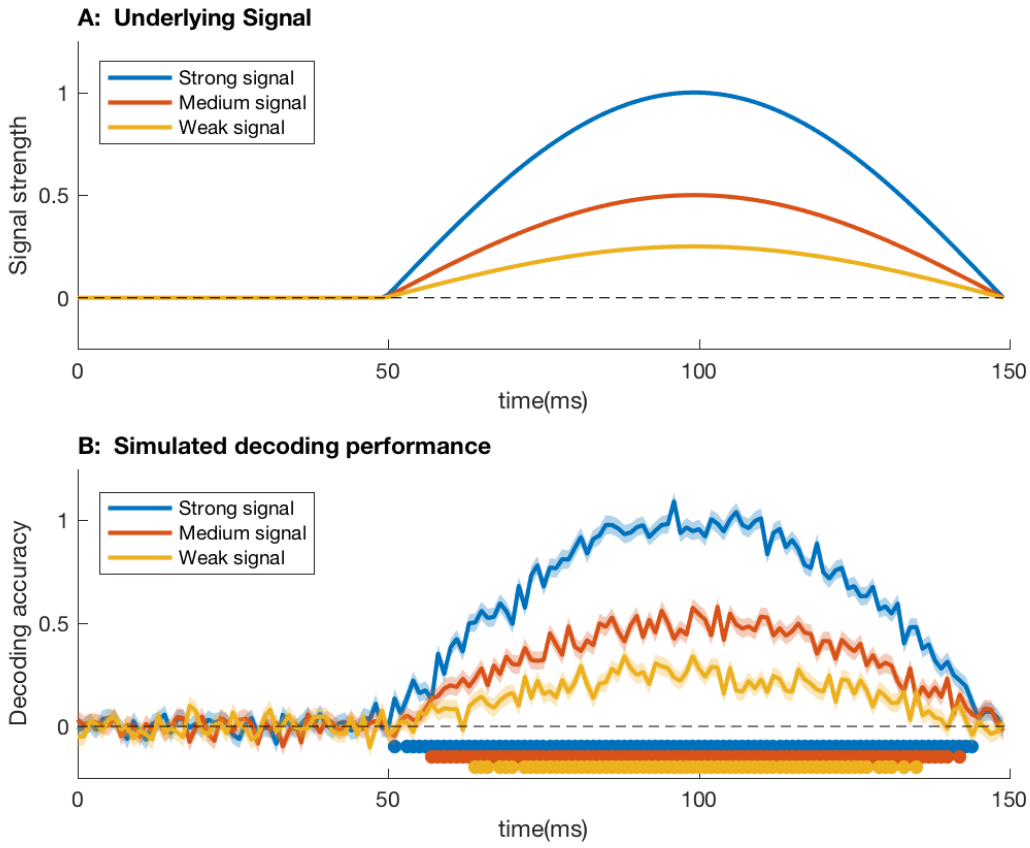

Figure 14. Demonstration of how the strength of peak decoding affects decoding onsets using stimulated data. A. Three data sets were simulated to have the same onset and peak decoding latencies, but different peak strengths. **B.** Gaussian noise was added to the underlying signals in each set (500 trials per set, σ=1) and significant decoding (above zero) was assessed across the time-course (signed-rank test, FDR corrected). Coloured discs above the x-axis indicate time points with significant decoding.

A second caveat specific to time-series decoding is that caution is required when interpreting (differences in) onsets of significant decoding. The time at which decoding is first significant for an experimental condition is determined by the underlying strength of the signal. For example, when the strength of peak decoding differs between two conditions (e.g., one is much easier to decode than the other), this will also affect the relative onset of decoding. This is illustrated in Figure 14. Three simulated data sets

were constructed to have the same decoding onset (50ms) and peak latency of decoding (100ms), but different signal strengths (see Figure 14A). To evaluate how signal strength influences decoding onset, Gaussian noise was added to each data set and significance testing was conducted to find the onset of decoding (signed-rank test across time points, FDR corrected). The outcome of the simulation is plotted in Figure 14B. Note that even though these simulated data sets were constructed to have an identical 'true' onset of decoding, the onset of significant decoding is earlier for the set with a strong signal and much later for the set with the weak signal. This underscores the ambiguity in interpreting onset differences: it cannot be assumed that an earlier decoding onset reflects a true onset difference in the availability of decodable information between conditions. (Isik et al., 2014) addresses this issue by using less data for the condition that had higher peak decoding, and by equalizing the peaks across conditions before determining decoding onset.

Third, as noted earlier, filtering the signal can smear out information over time. An extreme example (using a step function) is illustrated in Figure 15, using simulated data with a signal occurring at 50ms. To demonstrate the effect of filtering, Gaussian noise was added to the signal, and low-pass filters were applied with different cut-off frequencies using the *ft_preproc_lowpassfilter* function (using the default Butterworth 4^{th} order two-pass IIR filter) from the FieldTrip toolbox (Oostenveld et al., 2010). The result of lowering the cut-off frequency is increased signal distortion. Applying a 30Hz low-pass filter resulted in a signal that was significantly different from zero 40ms earlier in the time-series, compared to the simulated 'true' onset at 50ms. However, the effect is substantially reduced by applying much higher filter cut-offs, e.g., 200Hz. Therefore,

interpretations based on the timing of decoding signatures relative to the stimulus should be avoided when using filters with a low cut-off frequency (Vanrullen, 2011).

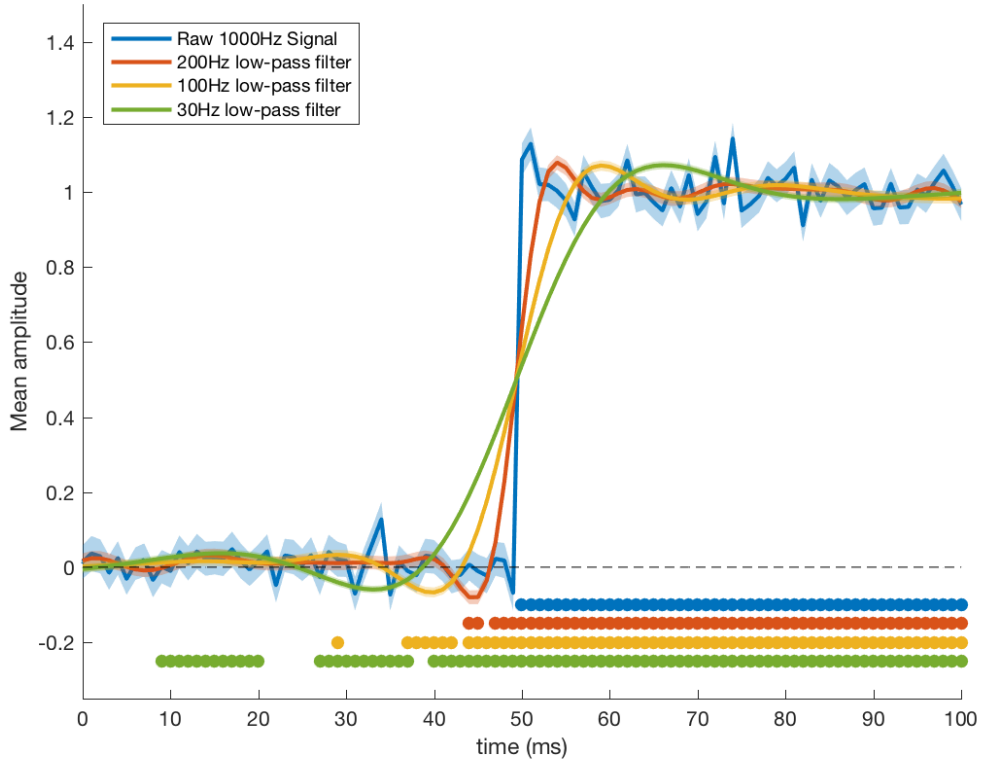

Figure 15. The effect of low-pass filtering on decoding onset. In this example, a signal with onset at 50ms was simulated with added Gaussian noise (500 trials, σ=1). The signal was then low-pass filtered using different cut-off frequencies. Time points where the trial average differed significantly from zero (signed-rank test, FDR corrected) are indicated by the coloured discs above the x-axis.

Finally, decoding studies require careful experimental design to avoid confounds in the classifier analysis. The considerations vital to designing decoding studies are not necessarily the same as that for univariate analysis. Accordingly, care must be taken when re-analysing data not originally intended for a decoding analysis. The high sensitivity of classifiers means that if there are any differences between classes other than the intended manipulations, it is likely that the classifier will exploit this

information, making it easy to introduce experimental confounds. An example is the effect of the subject's behavioural responses. In our example MEG experiment, the response buttons (to respond 'animate' and 'inanimate') were switched every block. If response mapping were uniform across blocks, response would be confounded with stimulus category, as a left button response would always correspond to 'animate', and right for 'inanimate'. The physical pressing of the button would generate corresponding brain signals, for example in motor areas, and this would provide a signal in the whole-brain MEG data that would correlate perfectly with the class conditions. In this case, it would be unclear whether the classifier decoded the intended experimental manipulation of 'animacy' or simply the subject's motor responses. Alternatively, a classifier may distinguish between two conditions or categories of stimuli based on a confounding factor that co-varies with class membership (e.g., differential attention to two conditions, leading to greater overall signal for one class) rather than the manipulation (e.g., difference in visual features or task difficulty) intended by the experimental design.

Further, even with carefully controlled designs, the interpretation of decoding studies must be executed with caution. Decoding studies may conclude that condition A is decodable from condition B; however, the source of decodable information usually remains elusive (Carlson & Wardle, 2015; Naselaris & Kay, 2015). One notable example of this is the current debate surrounding the source of orientation decoding in fMRI (e.g., (Alink, Krugliak, Walther, & Kriegeskorte, 2013; Carlson, 2014; Carlson & Wardle, 2015; Clifford & Mannion, 2015; Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Freeman, Brouwer, Heeger, & Merriam, 2011; Kamitani & Tong, 2005; Mannion, McDonald, & Clifford, 2009; Pratte, Sy, Swisher, & Tong, 2016). Despite a

decade of orientation decoding in early visual cortex with fMRI, it is still debated whether any information at the sub-voxel level (e.g., within-voxel biases in orientation-specific columnar responses) contributes to the decodable signal (op de Beeck, 2010). The interpretation of the source of decodable signals in neuroimaging remains one of the central challenges facing the application of MVPA techniques to advancing our understanding of information processing in the human brain (de-Wit et al., 2016).

Acknowledgements

This research was supported by an Australian Research Council (ARC) Future Fellowship (FT120100816) and ARC Discovery project (DP160101300) awarded to T.A.C. S.G.W. is supported by an Australian NHMRC Early Career Fellowship (APP1072245). The authors declare no competing financial interests.

References

Alink, A., Krugliak, A., Walther, A., & Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Frontiers in Psychology*, *4*, 493. https://doi.org/10.3389/fpsyg.2013.00493

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage*, *141*, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040

Allison, B. Z., Wolpaw, E. W., & Wolpaw, J. R. (2007). Brain–computer interface systems: progress and prospects. *Expert Review of Medical Devices*, *4*(4), 463–474. https://doi.org/10.1586/17434440.4.4.463

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bennett, C. M., Baird, A., Miller, M. B., & Wolford, G. L. (2011). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, *1*, 1–5.

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience*, *4*(4), 417–422. https://doi.org/10.1093/scan/nsp053

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4). springer New York.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage*, *56*(2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048

Bode, S., Sewell, D. K., Lilburn, S., Forte, J. D., Smith, P. L., & Stahl, J. (2012). Predicting Perceptual Decision Biases from Early Brain Activity. *Journal of Neuroscience*, *32*(36), 12488–12498. https://doi.org/10.1523/JNEUROSCI.1708-12.2012

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Carlson, T. A. (2014). Orientation Decoding in Human Visual Cortex: New Insights from an Unbiased Perspective. *The Journal of Neuroscience*, *34*(24), 8373–8383. https://doi.org/10.1523/JNEUROSCI.0548-14.2014

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, *11*(10), 9. https://doi.org/10.1167/11.10.9

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of Activity in the Categorical Representations of Objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. https://doi.org/10.1162/jocn.2003.15.5.704

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2013). The Emergence of Semantic Meaning in the Ventral Temporal Pathway. *Journal of Cognitive Neuroscience*, *26*(1), 120–131. https://doi.org/10.1162/jocn_a_00458

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1. https://doi.org/10.1167/13.10.1

Carlson, T. A., & Wardle, S. G. (2015). Sensible decoding. *NeuroImage*, *110*, 217–218. https://doi.org/10.1016/j.neuroimage.2015.02.009

Cauchoix, M., Arslan, A. B., Fize, D., & Serre, T. (2012). The Neural Dynamics of Visual Processing in Monkey Extrastriate Cortex: A Comparison between Univariate and Multivariate Techniques. In G. Langs, I. Rish, M. Grosse-Wentrup, & B. Murphy (Eds.), *Machine Learning and Interpretation in Neuroimaging* (pp. 164–171). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34713-9_21

Cauchoix, M., Barragan-Jason, G., Serre, T., & Barbeau, E. J. (2014). The Neural Dynamics of Face Detection in the Wild Revealed by MVPA. *The Journal of Neuroscience*, *34*(3), 846–854. https://doi.org/10.1523/JNEUROSCI.3030-13.2014

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2010). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, *54*(4), 3028–3039. https://doi.org/10.1016/j.neuroimage.2010.10.073

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, *44*(1), 62–70. https://doi.org/10.1016/j.neuroimage.2008.05.021

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204. https://doi.org/10.1016/j.neuroimage.2015.07.011

Clifford, C. W. G., & Mannion, D. J. (2015). Orientation decoding: Sense in spirals? *NeuroImage*, *110*, 219–222. https://doi.org/10.1016/j.neuroimage.2014.12.055

Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice*. MIT Press.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/10.1016/S1053-8119(03)00049-1

Curran, E. A., & Stokes, M. J. (2003). Learning to control brain activity: A review of the production and control of EEG components for driving brain–computer interface (BCI) systems. *Brain and Cognition*, *51*(3), 326–336. https://doi.org/10.1016/S0278-2626(03)00036-8

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*(1), 44–58. https://doi.org/10.1016/j.neuroimage.2008.06.037

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*(5), 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Downing, P. E., Chan, A. W.-Y., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain Specificity in Visual Cortex. *Cerebral Cortex*, *16*(10), 1453–1461. https://doi.org/10.1093/cercor/bhj086

Duncan, K. K., Hadjipapas, A., Li, S., Kourtzi, Z., Bagshaw, A., & Barnes, G. (2010). Identifying spatially overlapping local cortical networks with MEG. *Human Brain Mapping*, *31*(7), 1003–1016. https://doi.org/10.1002/hbm.20912

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321. https://doi.org/10.3758/BF03330618

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, *70*(6), 510–523. https://doi.org/10.1016/0013-4694(88)90149-6

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magnetic Resonance Imaging*, *26*(7), 921–934. https://doi.org/10.1016/j.mri.2008.01.052

Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation Decoding Depends on Maps, Not Columns. *The Journal of Neuroscience*, *31*(13), 4792–4804. https://doi.org/10.1523/JNEUROSCI.5160-10.2011

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981. https://doi.org/10.1038/nn.3402

Goddard, E., Carlson, T. A., Dermody, N., & Woolgar, A. (2016). Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, *128*, 385–397. https://doi.org/10.1016/j.neuroimage.2016.01.006

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Brain Imaging Methods*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Guimaraes, M. P., Wong, D. K., Uy, E. T., Grosenick, L., & Suppes, P. (2007). Single-Trial Classification of MEG Recordings. *IEEE Transactions on Biomedical Engineering*, *54*(3), 436–443. https://doi.org/10.1109/TBME.2006.888824

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. https://doi.org/10.1007/BF02512476

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., … Pollmann, S. (2009). PyMVPA: a unifying approach to the analysis of neuroscientific data. *Frontiers in Neuroinformatics*, *3*, 3. https://doi.org/10.3389/neuro.11.003.2009

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Hill, N. J., Lal, T. N., Schröder, M., Hinterberger, T., Widman, G., Elger, C. E., … Birbaumer, N. (2006). Classifying Event-Related Desynchronization in EEG, ECoG and MEG Signals. In K. Franke, K.-R. Müller, B. Nickolay, & R. Schäfer (Eds.), *Pattern Recognition* (pp. 404–413). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11861898_41

Hogendoorn, H., Verstraten, F. A. J., & Cavanagh, P. (2015). Strikingly rapid neural basis of motion-induced position shifts revealed by high temporal-resolution

EEG pattern classification. *Vision Research*, *113*, *Part A*, 1–10. https://doi.org/10.1016/j.visres.2015.05.005

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, *310*(5749), 863–866. https://doi.org/10.1126/science.1117593

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Jackson, J. E. (1991). *A user's guide to principal components* (Vol. 587). John Wiley & Sons.

Jafarpour, A., Horner, A. J., Fuentemilla, L., Penny, W. D., & Duzel, E. (2013). Decoding oscillatory representations and mechanisms in memory. *Neuropsychologia*, *51*(4), 772–780. https://doi.org/10.1016/j.neuropsychologia.2012.04.002

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology*, *115*(4), 2246–2250. https://doi.org/10.1152/jn.01074.2015

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

King, J.-R., Gramfort, A., Schurger, A., Naccache, L., & Dehaene, S. (2014). Two Distinct Dynamic Modes Subtend the Detection of Unexpected Sounds. *PLoS ONE*, *9*(1), e85791. https://doi.org/10.1371/journal.pone.0085791

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., & others. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism*, *30*(9), 1551–1557. https://doi.org/10.1038/jcbfm.2010.86

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., & Birbaumer, N. (2001). Brain–computer communication: Unlocking the locked in. *Psychological Bulletin*, *127*(3), 358–375. https://doi.org/10.1037/0033-2909.127.3.358

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. MIT press.

Mannion, D. J., McDonald, J. S., & Clifford, C. W. G. (2009). Discrimination of the local orientation structure of spiral Glass patterns early in human visual cortex. *NeuroImage*, *46*(2), 511–515. https://doi.org/10.1016/j.neuroimage.2009.01.052

Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage*, *67*, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027

Meyers, E. M. (2013). The neural decoding toolbox. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00008

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology*, *100*(3), 1407–1419. https://doi.org/10.1152/jn.90248.2008

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. https://doi.org/10.1016/j.neuroimage.2010.05.051

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Müller, K., Anderson, C. W., & Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*(2), 165–169. https://doi.org/10.1109/TNSRE.2003.814484

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, *167*(1), 82–90. https://doi.org/10.1016/j.jneumeth.2007.09.022

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. https://doi.org/10.1093/scan/nsn044

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., & Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, *117*(1), 12–22. https://doi.org/10.1016/j.bandl.2010.09.013

Naselaris, T., & Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends in Cognitive Sciences*, *19*(10), 551–554. https://doi.org/10.1016/j.tics.2015.07.005

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–815. https://doi.org/10.1016/j.neuroimage.2012.04.014

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Comput Biol*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., … Laureys, S. (2014). Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *NeuroImage: Clinical*, *4*, 687–694. https://doi.org/10.1016/j.nicl.2014.04.004

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, 27. https://doi.org/10.3389/fninf.2016.00027

op de Beeck, H. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047

Pantazis, D., Nichols, T. E., Baillet, S., & Leahy, R. M. (2005). A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *NeuroImage*, *25*(2), 383–394. https://doi.org/10.1016/j.neuroimage.2004.09.040

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1, Supplement 1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Pratte, M. S., Sy, J. L., Swisher, J. D., & Tong, F. (2016). Radial bias is not necessary for orientation decoding. *NeuroImage*, *127*, 23–33. https://doi.org/10.1016/j.neuroimage.2015.11.066

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/jocn_a_00924

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., & Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *The Journal of Neuroscience*, *33*(18), 7691–7699. https://doi.org/10.1523/JNEUROSCI.3905-12.2013

Redcay, E., & Carlson, T. A. (2015). Rapid neural discrimination of communicative gestures. *Social Cognitive and Affective Neuroscience*, *10*(4), 545–551. https://doi.org/10.1093/scan/nsu089

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLoS Comput Biol*, *11*(6), e1004316. https://doi.org/10.1371/journal.pcbi.1004316

Sandberg, K., Bahrami, B., Kanai, R., Barnes, G. R., Overgaard, M., & Rees, G. (2013). Early Visual Responses Predict Conscious Face Perception within and between Subjects during Binocular Rivalry. *Journal of Cognitive Neuroscience*, *25*(6), 969–985. https://doi.org/10.1162/jocn_a_00353

Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., & Desain, P. (2010). Name that tune: Decoding music from the listening brain. *NeuroImage*, *56*(2), 843–849. https://doi.org/10.1016/j.neuroimage.2010.05.084

Schreiber, K., & Krekelberg, B. (2013). The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging. *PLOS ONE*, *8*(7), e69328. https://doi.org/10.1371/journal.pone.0069328

Schwarzkopf, D. S., & Rees, G. (2011). Pattern classification using functional magnetic resonance imaging. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 568–579. https://doi.org/10.1002/wcs.141

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Simanova, I., van Gerven, M. A. J., Oostenveld, R., & Hagoort, P. (2014). Predicting the Semantic Category of Internally Generated Words from Neuromagnetic Recordings. *Journal of Cognitive Neuroscience*, 1–11. https://doi.org/10.1162/jocn_a_00690

Simanova, I., van Gerven, M., Oostenveld, R., & Hagoort, P. (2010). Identifying Object Categories from Event-Related EEG: Toward Decoding of Conceptual Representations. *PLoS ONE*, *5*(12), e14465. https://doi.org/10.1371/journal.pone.0014465

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, *62*(1), 451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., … Blankertz, B. (2012). Review of the BCI competition IV. *Neuroprosthetics*, *6*, 55. https://doi.org/10.3389/fnins.2012.00055

Thirion, B., Pedregosa, F., Eickenberg, M., & Varoquaux, G. (2015). Correlations of correlations are not reliable statistics: implications for multivariate pattern analysis. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*. Retrieved from https://hal.inria.fr/hal-01187297/

van de Nieuwenhuijzen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., & van Gerven, M. A. J. (2013). MEG-based decoding of the spatiotemporal

dynamics of visual category perception. *NeuroImage*, *83*, 1063–1073. https://doi.org/10.1016/j.neuroimage.2013.07.075

van Gerven, M. A. J., Maris, E., Sperling, M., Sharan, A., Litt, B., Anderson, C., … Jacobs, J. (2013). Decoding the memorization of individual stimuli with direct human brain recordings. *NeuroImage*, *70*, 223–232. https://doi.org/10.1016/j.neuroimage.2012.12.059

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, *44*(9), 867–880. https://doi.org/10.1109/10.623056

Vanrullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Perception Science*, *2*, 365. https://doi.org/10.3389/fpsyg.2011.00365

Vidal, J. J. (1973). Toward Direct Brain-Computer Communication. *Annual Review of Biophysics and Bioengineering*, *2*(1), 157–180. https://doi.org/10.1146/annurev.bb.02.060173.001105

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage*, *132*, 59–70. https://doi.org/10.1016/j.neuroimage.2016.02.019

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1(6)*, 80–83.

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*, 123. https://doi.org/10.3389/fnsys.2015.00123

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, *113*(6), 767–791. https://doi.org/10.1016/S1388-2457(02)00057-3

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *108*(21), 8850–8855. https://doi.org/10.1073/pnas.1100999108

Chapter 3

Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions

Tijl Grootswagers[1,2,3], J. Brendan Ritchie[4], Susan G. Wardle[1,2], Andrew Heathcote[5,6], & Thomas A. Carlson[2,3]

[1] Department of Cognitive Science, Macquarie University, Australia

[2] ARC Centre of Excellence in Cognition and its Disorders, Australia

[3] School of Psychology, University of Sydney, Australia

[4] KU Leuven, Belgium

[5] University of Tasmania, Australia

[6] University of Newcastle, Australia

Note: This chapter is currently under review at the *Journal of Cognitive Neuroscience*

Abstract

Animacy is a robust organizing principle amongst object category representations in the human brain. Using multivariate pattern analysis methods (MVPA), it has been shown that distance to the decision boundary of a classifier trained to discriminate neural activation patterns for animate and inanimate objects correlates with observer reaction times for the same animacy categorization task (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Ritchie, Tovar, & Carlson, 2015). Using MEG decoding, we tested if the same relationship holds when a stimulus manipulation (degradation) increases task difficulty, which we predicted would systematically decrease the distance of activation patterns from the decision boundary, and increase reaction times. In addition, we tested whether distance to the classifier boundary correlates with drift rates in the Linear Ballistic Accumulator (Brown & Heathcote, 2008). We found that distance to the classifier boundary correlated with reaction time, accuracy, and drift rates in an animacy categorization task. Split by animacy, the correlations between brain and behaviour were sustained for longer over the time course for animate than for inanimate stimuli. Interestingly, when examining the distance to the classifier boundary during the peak correlation between brain and behaviour, we found that only degraded versions of animate, but not inanimate, objects had systematically shifted towards the classifier decision boundary as predicted. Our results support an asymmetry in the representation of animate and inanimate object categories in the human brain.

1 Introduction

Object recognition is a fast, reliable, and effortless process for humans. Early visual areas in the brain respond to simple visual features (e.g., edges, luminance contrast, or orientation), and further along the ventral stream sensitivity to objects and object categories (e.g., faces, animals, or tools) emerges (Grill-Spector & Weiner, 2014). Using multivariate pattern analysis (MVPA), several studies have analysed pattern similarities between the neural representation of objects in inferior temporal cortex (ITC) to study its representational structure, with both fMRI (Edelman, Grill-Spector, Kushnir, & Malach, 1998; Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013), and MEG (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014, 2016). These studies have provided evidence of a categorical organization in ITC, following the observation that objects belonging to the same category tend to evoke similar patterns of neural activation (Connolly et al., 2012; Haxby et al., 2001; Kriegeskorte et al., 2008; Sha et al., 2015). One robust categorical structure is the animate/inanimate distinction in human and primate ITC (Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008). Animate stimuli (e.g., humans or animals) evoke brain activation patterns that are more similar to other animate exemplars than to inanimate stimuli (e.g., plants, tools, vehicles) (Carlson et al., 2013; Cichy et al., 2014; Downing, Jiang, Shuman, & Kanwisher, 2001; Kiani et al., 2007; Kriegeskorte et al., 2008).

A current topic of debate is whether categorical information in the brain, as revealed using MVPA decoding, is read-out in behaviour (cf. de-Wit, Alexander, Ekroll, & Wagemans, 2016; Williams, Dang, & Kanwisher, 2007). For the presence of decodable

information in neuroimaging activation patterns related to a stimulus or task does not necessarily entail that this information underlies related behaviour. One recent approach to linking activation spaces to behaviour (Ritchie & Carlson, 2016) is inspired by distance-to-bound models of reaction time (Ashby & Maddox, 1994; Pike, 1973). According to distance-to-bound models, evidence close to a decision boundary is more ambiguous, reflecting greater difficulty in categorization, while evidence far from the decision boundary is less ambiguous with regards to category membership. Assuming that response time is a function of stimulus discriminability, and that a classifier decision boundary (which is used in MVPA decoding) reflects an observer's decision boundary, then reaction times should negatively correlate with distance from the boundary; for example, stimuli that are faster to categorize should be neurally represented as further from the classifier decision boundary (Ritchie & Carlson, 2016).

In a previous application of the RT-distance approach to MEG (Ritchie et al., 2015), a linear discriminant classifier (LDA) was trained to discriminate the MEG channel activation for animate from inanimate stimuli. Next, the distance of each stimulus pattern to the classifier boundary in high-dimensional space were rank-order correlated to human reaction times for categorizing the same stimuli as animate/inanimate. As predicted, the distance to boundary negatively correlated with reaction time. Moreover, the correlation over time tracked the MEG decoding time-series (Ritchie et al., 2015). Interestingly, when analysing the animate and inanimate stimuli separately, an asymmetry was observed: the correlation between distance to boundary and reaction time was driven by the animate stimuli (Carlson et al., 2014; Ritchie et al., 2015).

A good test of the RT-distance hypothesis is to manipulate task difficulty experimentally, and the effects of this behavioural manipulation on representational space. The RT-distance hypothesis has so far been tested on the differences *between* objects. For example, in Carlson et al, (2014), an ostrich was closer to the animacy decision boundary than a human face, and participants were slower to categorize the ostrich as 'animate'. The effect of increasing categorization difficulty of single object images on the distance to the classifier boundary in activation space has not yet been tested. According to the RT-distance hypothesis, a change in behaviour resulting from manipulating categorization task difficulty should be matched by a corresponding shift of the stimulus set in representational space. Numerous studies have shown that degrading object stimuli reduces categorization performance, such as by scrambling the image phase (e.g., Philiastides, Ratcliff, & Sajda, 2006; Philiastides & Sajda, 2006; Wichmann, Braun, & Gegenfurtner, 2006), scrambling image amplitude (e.g., Gaspar & Rousselet, 2009), reducing luminance contrast (e.g., Macé, Delorme, Richard, & Fabre-Thorpe, 2009; Macé, Thorpe, & Fabre-Thorpe, 2005), or blurring the image (e.g., Bruner & Potter, 1964; Párraga, Troscianko, & Tolhurst, 2000, 2005; Wyatt & Campbell, 1951). For degraded stimuli with longer categorization RTs, the RT-distance hypothesis predicts that these stimuli will be located closer to the animacy decision boundary, producing a 'compression' of representational space compared to that for the original versions of the stimuli (Fig 1A). This predicts a correlation between the shorter distance to boundary and slower accumulation rates (Fig 1B). Note that even in the case of inanimate stimuli where there was no correlation (Ritchie et al., 2015), one would still expect a correlation when including both clear and degraded versions because of the compression of the general representational space (Fig 1B).

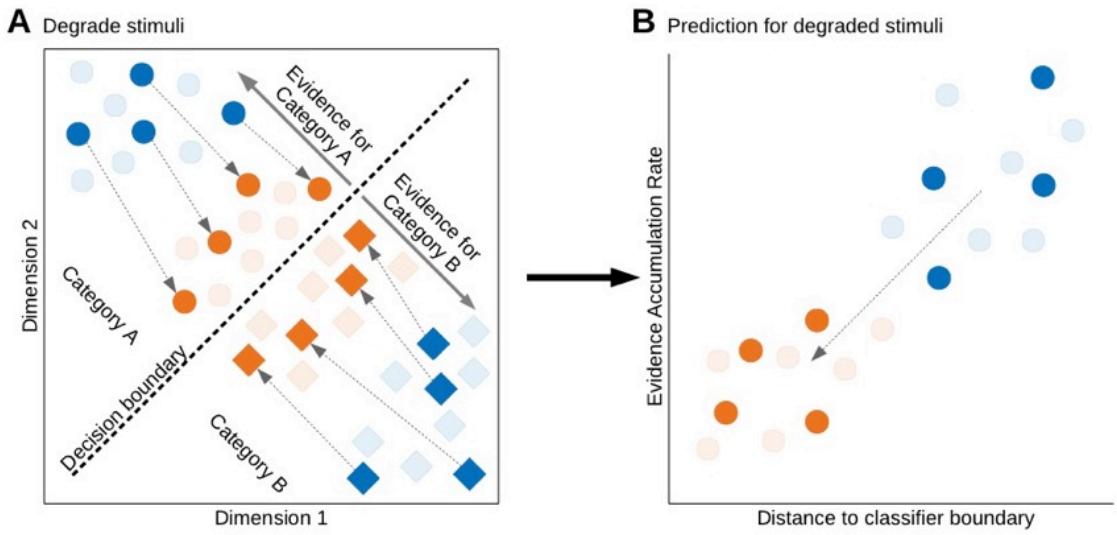

Figure 1. The predicted effect of degrading stimuli on their location in representational space. A. Stimuli of two categories are illustrated as circles (animate objects) and squares (inanimate objects) in representational space (only two dimensions are plotted here for visualization). The blue shapes represent the stimuli in a clear state, and the orange are their degraded counterparts. If distance to a classifier boundary is taken as representing evidence for a decision, this predicts that the degraded (orange) versions of the stimuli will be located closer to the classifier decision boundary (dashed line) that separates the stimulus categories than the clear (blue) versions. **B.** Stimuli from one category (animate objects) and their distance to the classifier boundary versus their rate of evidence accumulation for an animacy categorization decision. The RT-distance hypothesis predicts that the degraded stimuli, which have moved closer to the decision boundary, also have slower evidence accumulation rates (and therefore longer RTs). Note that the same prediction holds for the other category (inanimate objects).

The dominant models of reaction time appeal to some form of evidence accumulation process (Ashby, 2000; Ratcliff, 1985); That is, evidence for a decision (e.g., animate or inanimate) accumulates over time, and the response is made when the amount of evidence reaches a certain threshold (Brown & Heathcote, 2008; Gold & Shadlen, 2007; Ratcliff & Rouder, 1998). Distance to the classifier boundary can be linked to evidence accumulation. Carlson et al., (2014) simulated evidence accumulation with a sequential

analysis model, using the distance to the boundary for each object exemplar as a proxy for evidence strength. They found that accumulation rate correlated with categorization RTs, providing support for the theoretical link between distance to boundary and evidence accumulation (Carlson et al., 2014; Ritchie et al., 2015). One way to build upon this would be to more closely relate distance to evidence accumulation; beyond correlating distance with median RTs, an existing model of evidence accumulation can be fit to subjects' behavioural data, yielding independent model parameter estimates that can be correlated with distance to the boundary. As an accumulator model provides a more complete characterization of categorization behaviour than average reaction times, it may provide a better measure to correlate to representational distances.

The aim of the present study was two-fold. First, using MEG decoding we sought to test the prediction that degrading object exemplar would compress the representational space, which would correlate with slower animacy RTs compared to un-degraded, or clear versions of the same stimuli (Figure 1). Secondly, we aimed to test the RT-distance hypothesis in the context of an existing model of RT distributions and choice accuracy, the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008), in order to evaluate whether distance to boundary can be related more directly to evidence-accumulation model parameters.

2 Methods

2.1 Participants

All participants gave informed consent in writing prior to the experiment. The study was conducted with the approval of the Macquarie University Human Research Ethics Committee. 100 participants were recruited on Amazons Mechanical Turk (MTurk) to determine the level of stimulus degradation needed for equal object recognition performance across all stimuli (see Section 2.2). For the second part of the study, 20 healthy volunteers (4 males; mean age 29.3 years) with normal or corrected-to-normal vision participated in the MEG experiment. Participants in both experiments were financially compensated for their time. All analysis procedures were performed in Matlab, using the statistics and machine learning toolbox.

2.2 Stimuli

We constructed a set of 48 visual object stimuli including 24 animals and 24 inanimate objects (natural and man-made) on a phase scrambled natural image background in both a clear and a degraded condition. First, high resolution images (> 512 x 512 pixels) of various objects were collected via an internet search. We selected images that showed prototypical viewpoints of objects. In order to test the RT-distance hypothesis more generally, no human or human face images were included in the stimulus set as they are generally outliers both behaviourally and in the brain's response compared to other object exemplars (i.e., face stimuli tend to have fast categorization RTs and produce pronounced responses in neuroimaging data). It is therefore possible that the inclusion

of face stimuli could disproportionately explain any observed correlations between RT and neural distance. All exemplar images used in the study are shown in Figure 2A. As colour can be a salient cue for image recognition (Biederman & Ju, 1988; Joseph & Proffitt, 1996; Ostergaard & Davidoff, 1985; Wurm, Legge, Isenberg, & Luebker, 1993) and would make the degrading process (described hereafter) less effective, grey-scaled versions of the object images were used in the experiment. A different random noise background was created for each exemplar by phase scrambling a natural image of a forest scene (see Figure 2B). The object images were overlaid on the noise background, producing the stimuli for the clear condition (Figure 2C).

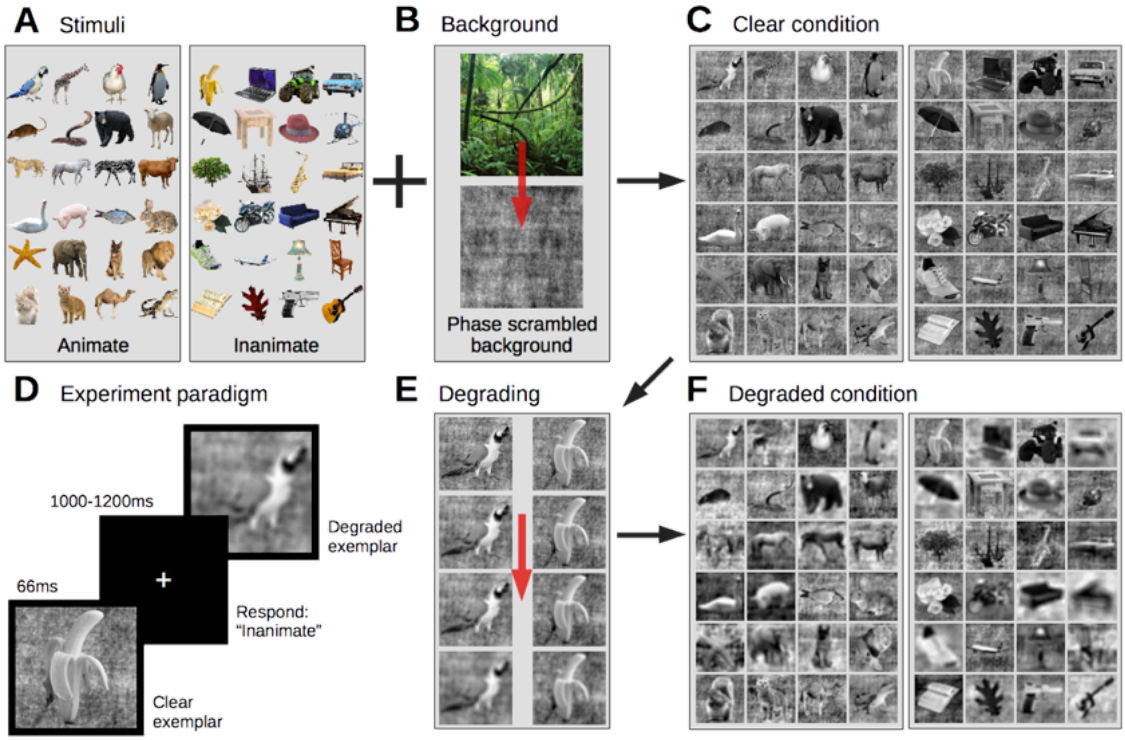

Figure 2. Stimuli and experiment design. A. Stimuli consisted of 24 animate and 24 inanimate objects. **B.** Stimuli were placed on a phase scrambled natural image background**. C.** All stimuli in the clear condition. **D.** In the MEG experiment, participants saw clear and degraded stimuli for 66ms in randomized order with a varying inter stimulus interval and were asked to report the stimulus animacy with a button press. **E.** To create the degraded condition, stimuli were gradually blurred by simulating defocus to a level where they were equally recognizable. **F.** All stimuli in the degraded condition.

We created a degraded condition by blurring the same set of images (including their background). As different objects require different levels of blur to equate recognition performance, we first measured the amount of blur required to impair recognition for each image. Our aim was for subjects to be able to perform the task given unlimited exposure time (i.e., correctly recognize the object), but to reduce their categorization performance under brief presentation duration (i.e., by reducing speed and/or accuracy). We simulated defocused blur using an image filter (Figure 3) that convolved the amplitude spectrum of the image in the Fourier domain (Figure 3A-B) with a Fourier-transformed cylinder function (a sombrero function, Figure 3C-D), where increasing the radius of the cylinder function results in a greater magnitude (Figure 3E) of defocus blur (Sonka, Hlavác, & Boyle, 2008). Images were then gradually degraded by increasing the radius in steps of 2 pixels, from a radius of one pixel (no degrading) to 59 (very degraded). The sequence of the image coming into focus was presented to the MTurk participants. Each participant saw all 48 stimuli from both animate and inanimate categories once starting from the most degraded state, while its level of focus was gradually increased (in steps of 2 pixel radius). Participants were instructed to press the spacebar as soon as they recognized the object in the picture. The stimulus was then removed from the screen, and participants entered a name for the stimulus. To check for correct recognition, the responses on the naming task were assessed manually for validity, to allow for variations in spelling or for synonymous names. For each exemplar, the amount of focus needed for 25% of the MTurk participants to correctly recognize (i.e., name) the object (see Figure 2F) was used as the blur filter parameters for that exemplar in the degraded condition. On average, a radius of 17 pixels (sd=4.68) was used for the animate exemplars and 20.5 pixels (sd=10.12) for the inanimate exemplars.

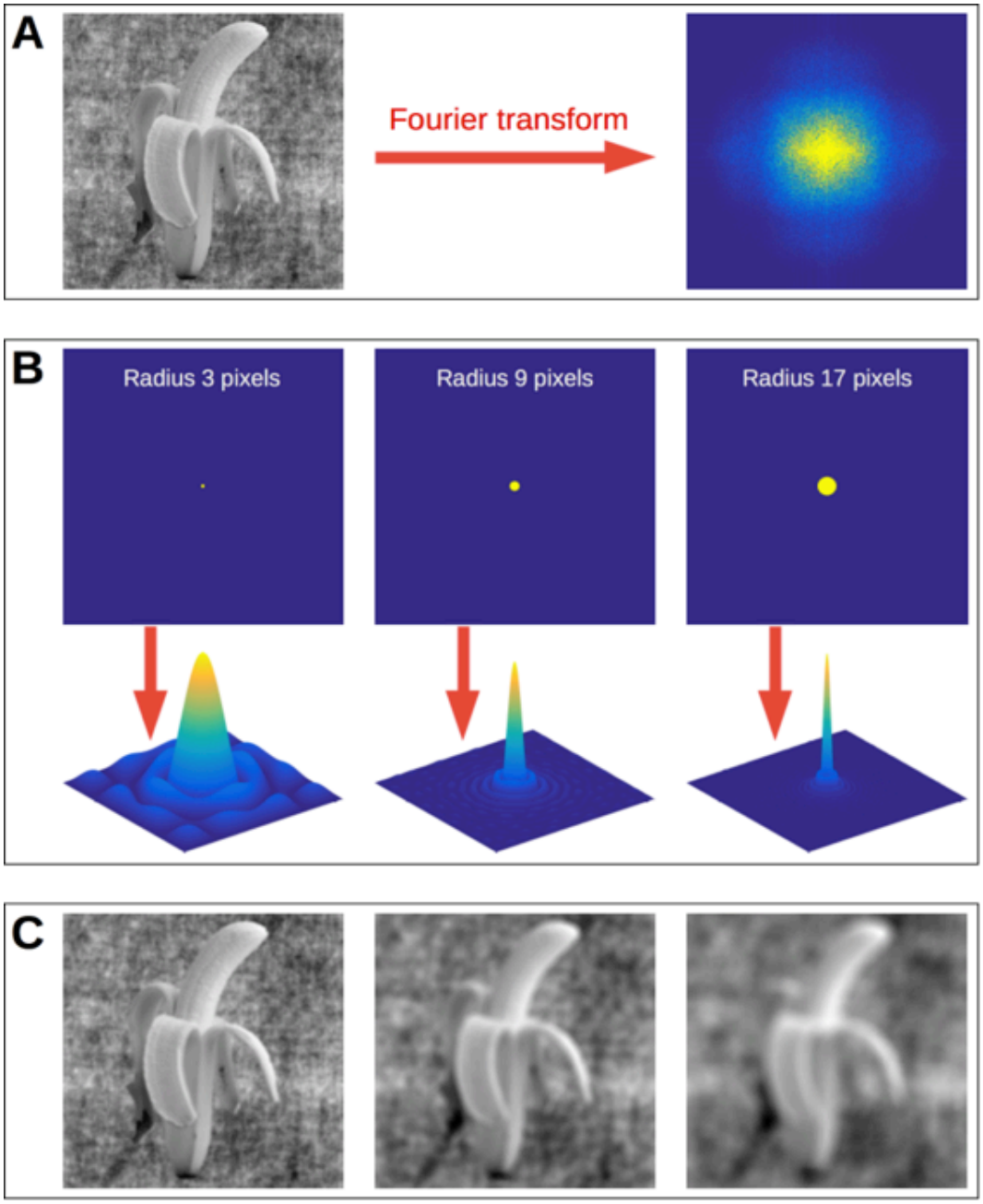

Figure 3. Image filter used to blur the experimental stimuli. A. The original image and its amplitude spectrum, which shows the typical energy pattern for natural images (high energy at low spatial frequencies (centre), and low energy at high spatial frequencies Field, 1987). **B.** Cylinder functions with radii of 3, 9, and 17 pixels and their Fourier transformed versions (sombrero functions), which are used as the image filters. **C.** Convolving the amplitude spectrum of the original image with the sombrero functions in the Fourier domain results in images with different levels of defocus blur by removing a significant proportion of energy at high spatial frequencies.

2.3 MEG Experiment Design

Before the MEG experiment, we confirmed that each participant could recognize all exemplars given unlimited presentation time, even in the degraded state, as the intention of the design was to decrease categorization speed and accuracy without making the stimuli unrecognizable. Stimuli in the degraded state were shown and participants were asked to name the object in each picture. If the participant failed to correctly recognize the stimulus, it was shown in the clear state, to ensure that all objects were correctly recognized. Next, the participant was trained on the task (outside the MEG), as the brief presentation duration and fast-pace of the categorization task required practice to master. If performance (animacy categorization accuracy) in the first block was lower than 80%, the participant was shown all the stimuli again in both states, to identify degraded exemplars they were unable to recognize, and then practiced again on a second training block. All 20 participants performed above 80% correct after the familiarization step.

Following the practice task, participants completed the MEG experiment. On each trial within a block, stimuli were projected (at 9x9° visual angle) on a black background for 66ms, followed by a fixation cross for a random duration between 1000 and 1200ms. Participants were asked to categorize the stimulus as animate or inanimate as fast and accurate as possible, using a button press (see also Figure 2D). The mapping of the response buttons alternated between blocks (Grootswagers, Wardle, & Carlson, 2017; Ritchie et al., 2015) and participants received feedback for the first 10 trials of a block (red or green cross), to ensure accurate mapping of the buttons. This was included to remove a class-specific motor preparation from the average signal and was chosen over

a single-trial randomized mapping to allow for fast-paced trials. Within each block, four repetitions of each exemplar in both conditions were presented in randomized order. Block duration was ~7 minutes. After each block, participant received feedback on performance (mean accuracy and number of missed trials). Each participant completed eight blocks, resulting in 32 trials per exemplar, 768 trials per category and condition (animate/inanimate, clear/degraded), and 3072 trials total per participant. The total time in the scanner was about one hour, including breaks between blocks.

2.4 MEG acquisition and preprocessing

Participants were fitted with a cap with 5 marker coils to track head movement during the session. The MEG signal was continuously sampled at 1000Hz from 160 axial gradiometers using a whole-head MEG system (Model PQ1160R-N2, KIT, Kanazawa, Japan) while participants lay in a supine position inside a magnetically shielded room (Fujihara Co. Ltd., Tokyo, Japan). Recordings were filtered online between 0.03Hz and 200Hz. We examined the delay in stimulus onsets (Ramkumar, Jas, Pannasch, Hari, & Parkkonen, 2013) by comparing the photodiode responses with the stimulus onset triggers (sent by the experiment script), and found a highly consistent delay of 56.26ms, for which we conservatively corrected by shifting the onset triggers back by 56ms. Recordings were sliced into 700ms epochs (-100 – 600ms post-stimulus onset). The trials were downsampled to 200Hz (5ms resolution) and transformed using principal component analysis (PCA), where the components that accounted for 99% of the variance were retained to reduce the dimensionality of the data (mean 62.25 components, sd 12.33). Finally, to increase signal to noise, 4 trials of each exemplar (with balanced response mappings) were averaged into pseudo-trials (Grootswagers et

al., 2017; Isik, Meyers, Leibo, & Poggio, 2014), leaving 8 pseudo-trials per exemplar in both conditions.

2.5 Sliding time window decoding

To investigate the decoding performance of animacy over time, sliding window Naïve Bayes classifiers were used on the pseudo-trials. To assess the difference in decoding performance between the clear and degraded conditions, three separate classifiers were used for decoding; one for each condition, and one for both conditions combined. At each 5ms interval *t*, the classifiers were trained and tested on a 25ms window (from *t*-25ms to *t*). The classifier performance was examined using leave-one-exemplar-out cross-validation (Carlson et al., 2013). In this method, the classifier is trained on the animacy of all-but-one exemplar, and tested on trials of the left out exemplar. This is repeated for each exemplar, and the mean decoding accuracy on the left out exemplars is used to assess generalization accuracy. Using this cross-validation method, the classifier has to generalize the *concept* of animacy, and cannot benefit from exploiting individual stimulus properties because the test exemplar is not in the training set (Carlson et al., 2013). We report the subject-averaged classifier accuracy over time, with significant above-chance accuracies assessed using a non-parametric Wilcoxon signed rank test. The False Discovery Rate (FDR) was used to control for false positives resulting from multiple comparisons.

2.6 Fitting LBA on individual subject behaviour

One of the aims of this study was to use a more complex model of evidence accumulation, and to then correlate distance to boundary with the drift rate parameters of the model, as well as RT and accuracy. The Linear Ballistic Accumulator is a mathematically tractable yet simple and complete model of evidence accumulation (LBA; Brown & Heathcote, 2008). LBA simultaneously takes into account both the full distribution of RTs (i.e., the mean, variability and positive skew of the RT distribution, see Luce, 1986) and task accuracy. LBA differs from other well-known accumulation models (e.g., the diffusion decision model Ratcliff, 1978, or the leaky competitive accumulator, Usher & McClelland, 2001) by modelling the response alternatives (here: animate and inanimate) with separate independent accumulators that accrue evidence accumulation in a linear and ballistic manner. In particular, the LBA model assumes that between-trial variability in the amount of evidence required for a decision, and the rate at which it accumulates, dominate over any moment to moment variability in evidence during a trial, whereas the latter source of noise plays a greater role in alternative models. These properties make LBA analytically simple and relatively easy to apply, which is ideal for this study, where each stimulus is treated as a separate condition, meaning that there are many parameters to estimate, each based on relatively few data points. To simplify the fitting procedure, the stimuli for both categories were separated into 6 bins based on accuracy. Next, a set of progressively more complex parameterizations were fit stepwise using maximum likelihood estimation (Donkin, Brown, & Heathcote, 2011; Rae, Heathcote, Donkin, Averell, & Brown, 2014). The most complex model parameterization, which was used here for further analysis, included clear/degraded, animate/inanimate, and stimulus as factors. Hence, each stimulus in both clear and degraded conditions had

a separate drift rate, which is important to note, as the goal was to correlate stimulus-specific distances with drift rates. To assess the reliability of the model fits, we compared the model predictions to the full distributions of accuracy and RT. Figure 4 shows that the LBA model provided a very accurate model of the effect of stimulus degradation and of differences between animate and inanimate images, both in terms of accuracy (Figure 4A) and the entire distribution of RTs (Figure 4B), and so provides a useful characterization of the participants' behaviour.

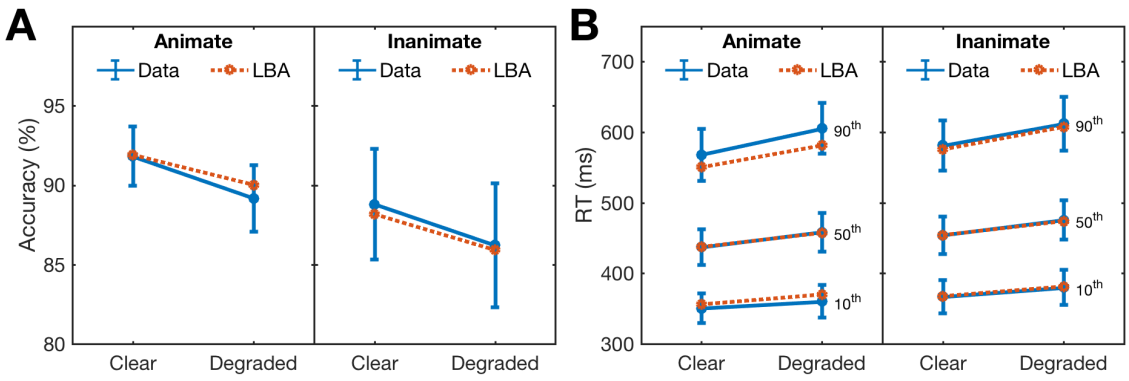

Figure 4. Fits of the LBA model to (A) accuracy and (B) the distribution of RT, with 95% confidence intervals. The RT distribution is illustrated by plotting the 90th percentile (upper lines, representing the slowest RTs), the median (50th percentile, middle lines) and the 10th percentile (lower lines, representing the fastest RTs).

2.7 Predicting behaviour from representational distance

Individual exemplars can be represented as points in representational space (i.e., a multidimensional feature space). To decode animacy, we applied a discriminant classifier (Gaussian Naïve Bayes) to optimize a decision boundary in multidimensional space, separating the neural patterns for animate and inanimate exemplars. According to the RT-distance hypothesis, the neural representations of the exemplars that are close (in multidimensional space) to this boundary are predicted to be more difficult to discriminate, as there is less evidence for the decision. This forms a prediction for behaviour, where less evidence for a decision would result in slower RTs or lower accuracies (Carlson et al., 2014; Ritchie & Carlson, 2016; Ritchie et al., 2015) , and, in the case of the LBA decision model, slower drift rates. For the current study, we predicted that degraded stimuli would be closer to the decision boundary in representational space, and therefore correlate with lower accuracy, slower reaction times, and slower drift rates.

We tested this prediction by repeating the following process at each time window: First, all the trials for each exemplar were averaged to create an average representation in multidimensional space for each exemplar (one for the exemplar in clear, and one for degraded state). In this space, a decision boundary for animacy was fitted (i.e., training a Naïve Bayes classifier), and the representational distance to this boundary for each exemplar was computed. This resulted in a distance-value for each exemplar in both clear and degraded states (24 animate and 24 inanimate exemplars in 2 states = 92 distance values). Next, the exemplar distances were rank-order correlated (using Spearman's ρ) to the two behavioural measures (median reaction time, mean accuracy) and the mean drift rate for the exemplars.

Repeating this process over time and subjects resulted in three time-varying correlations (two for each behavioural measure, and one for drift rate) for each subject. We then report the subject-averaged time-varying correlations, and used a non-parametric Wilcoxon signed rank test at each time window to test for significant above-zero correlations at the group level. Note that this approach differs from Ritchie et al., (2015), where distances and reaction times were averaged over subjects first, and the correlations were performed at the group level. False discovery rate (FDR) adjustment was used to control for false positives resulting from multiple comparisons. This process allowed us to compare the distance to boundary correlations with the three variables. We also sought to assess the possibility of modulating effects of animacy, for example whether the animate-inanimate asymmetry reported in Carlson et al., (2014) and Ritchie et al. (2015) was replicated, and whether it was affected by degrading the stimuli. For this, we computed the time-varying correlations separately for animate, and inanimate exemplars.

3 Results

3.1 Behavioural results

Overall, subjects performed well on the categorization task (mean accuracy 87.9%, sd 12.7) with a median RT of 457.5ms (sd 64.6). Figure 5 shows the median RT (Figure 5A) and mean accuracy (Figure 5B) separate for category (animate vs. inanimate) and stimulus condition (clear vs degraded). Analysis of variance showed that animate exemplars were easier to categorize than inanimate exemplars (compare left versus right

groups in each plot), with significantly faster median RTs ($F(1,19)=32.09$, $p<.0001$) and higher accuracies ($F(1,19)=6.27$, $p<.0001$). As predicted, degrading the stimuli made the animacy categorization task more difficult (compare blue and yellow lines), significantly lowering accuracy ($F(1,19)=41.81$, $p<.0001$) and increasing median RT ($F(1,19)=85.06$, $p<.0001$) for both animate and inanimate exemplars. There was no significant interaction between animacy and degrading for either median RT ($F(1,19)=0.04$, $p=0.84$), or accuracy ($F(1,19)=0.01$, $p=0.92$).

In order to test whether distance is related to evidence accumulation, we obtained drift rates for each exemplar and subject from the LBA fits. The LBA produces a separate drift rate parameter for the correct accumulator (i.e., the animate accumulator for animate stimuli, and the inanimate accumulator for inanimate stimuli) and the incorrect accumulator (i.e., the inanimate accumulator for animate stimuli and vice versa), however only the drift rate for the correct response accumulators were included in our analysis. The resulting drift rate parameters are summarized in Figure 5C, split by animacy and degradation. Degrading the stimuli resulted in significantly slower drift rates ($F(1,19)=41.20$, $p<.0001$). There was a main effect of animacy on drift rate ($F(1,19)=4.44$, $p<0.05$), and a significant interaction between animacy and stimulus clarity on drift rate ($F(1,19)=18.52$, $p<0.0001$), as degrading reduced the drift rate for animate exemplars more than inanimate exemplars (see Figure 5C). In sum, we confirmed that the degraded exemplars were generally harder to categorize (in terms of longer RT, lower accuracy, and slower drift rate), demonstrating our stimulus categorization difficulty manipulation (blur) was successful.

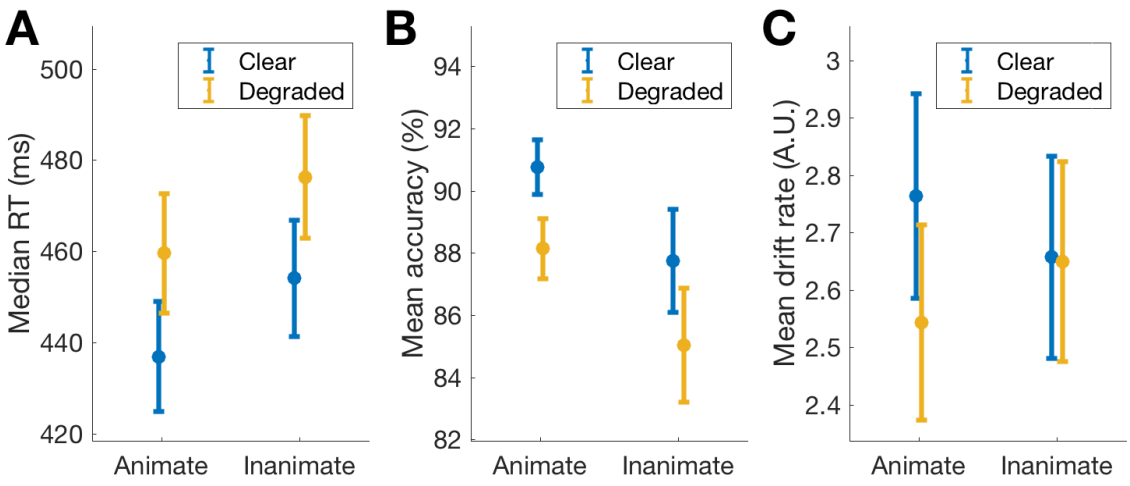

Figure 5. Behavioural results. The distributions of (**A**) the median reaction time, (**B**) accuracy, and (**C**) drift rate for the correct response accumulator, split up by exemplar category (animate/inanimate) and condition (clear/degraded, blue and orange lines). Error bars represent +/- SEM.

3.2 MEG Decoding

A prediction of degrading the stimuli is that decoding performance will be lower relative to that for clear stimuli. To evaluate the effect of blurring the stimuli on decoding performance over time, sliding time window classifiers were trained on predicting stimulus animacy at each time point. The results are presented in Figure 6, as mean cross-validated decoding accuracy over subjects, separately for clear and degraded stimuli. The decoding onset in the clear condition is at 75ms, which is consistent with previous findings showing an animacy decoding onset of approximately 60 - 80 ms (Carlson et al., 2013; Ritchie et al., 2015). In the degraded condition, decoding onset was 170 ms, and decoding performance was significantly lower than in the clear condition over the entire time course (black marks above the x-axis). Peak decoding performance for degraded objects was also later, 380 ms compared to 345 ms for clear stimuli. Initially, decoding performance for the combined data for clear and degraded

conditions closely matched that for the degraded condition, but then raised to a level closer to the performance in the clear condition. In sum, the neural patterns for blurred stimuli were more difficult to decode from the whole-brain MEG activation patterns than those for the clear versions of the same stimuli.

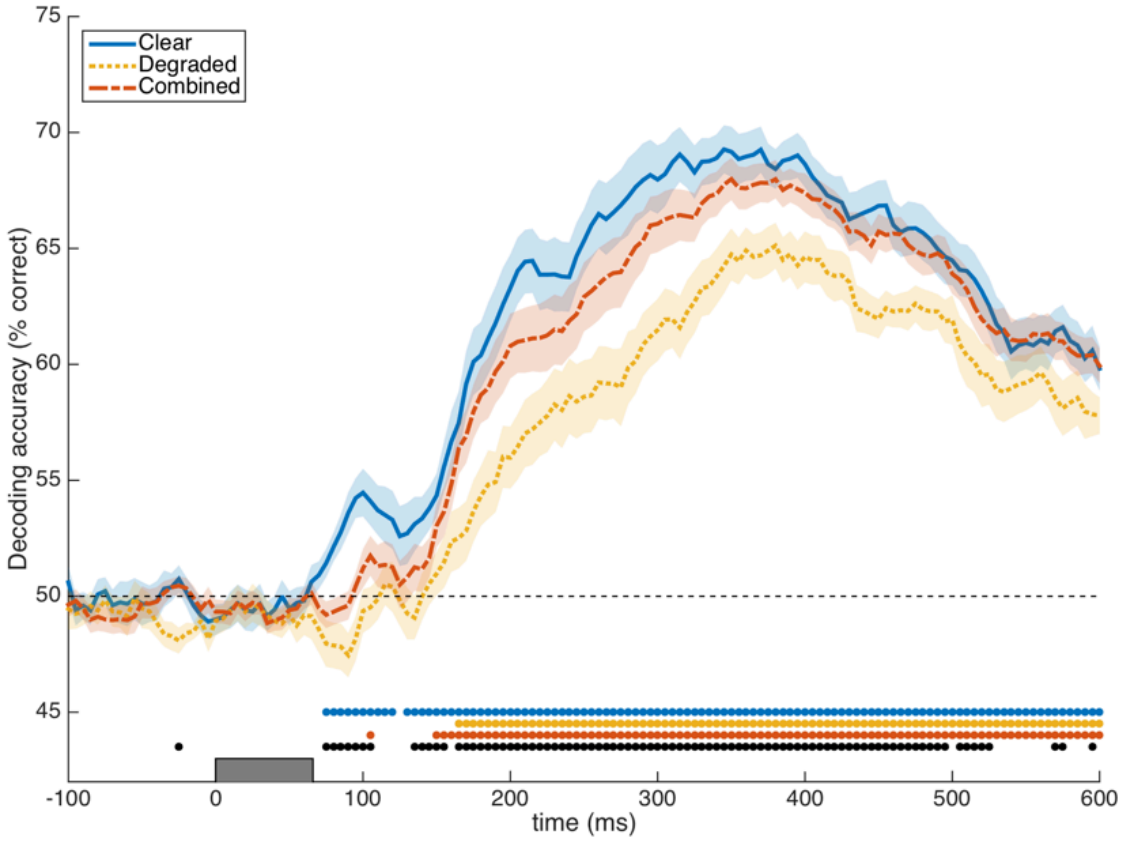

Figure 6. Decoding animacy from the MEG signal. Decoding was performed using leave-one-exemplar-out cross-validation with 25ms sliding time window classifiers. At each time point *t*, the graph shows the mean classifier accuracy over subjects at the window [*t*-25ms, *t*]. Shaded areas show standard error between subjects. Coloured marks above the x-axis indicate significant above-chance (50%) decoding. Black marks indicate significant (FDR-adjusted $p<0.05$) differences in classifier accuracy between the clear and degraded conditions. The grey bar on the x-axis indicates the time that the stimulus was on the screen (0-66ms).

3.3 Predicting behaviour from representational distance

To investigate the relationship between decodability and behaviour, we computed the time-varying distance to the classifier decision boundary for all the exemplars for each subject. These subject-specific distances were then rank-order correlated to the subject's behavioural measures for each exemplar: RT, accuracy, and the drift rate as fitted by LBA. Note that for accuracy and drift rate, the RT-distance hypothesis predicts a *positive* correlation (closer to the boundary corresponds to lower accuracy and slower drift rate), however a *negative* correlation is predicted between RT and distance (closer to the boundary corresponds to longer reaction times). For ease of comparison the sign of the correlation for RT was inverted in Fig 7.

Figure 7A shows the mean time-varying correlations over subjects for drift rate (green line), RT (red line), and accuracy (purple line). The peaks of the time-varying correlations are shown in the inset bar graphs. Although drift rate appears to have an earlier and higher peak correlation than RT and accuracy, this difference was not statistically significant. The similarity between these time-varying correlations is likely due to correlations among the different behavioural measures (e.g., exemplars with fast RTs are likely to have high accuracies). These results further show that accuracy and drift rate can be predicted equally well using distance to boundary. While the LBA model has been used before to fit non-human primate neural activations (Cassey, Heathcote, & Brown, 2014), this is the first time it has been related to neuroimaging data measured with MEG. Thus, our results are promising considering that drift rate more closely represents the accumulation of evidence for a decision, compared to RT or accuracy. In

sum, distance to boundary correlates with the behavioural measures as well as the fitted drift rate parameters and these follow similar trajectories.

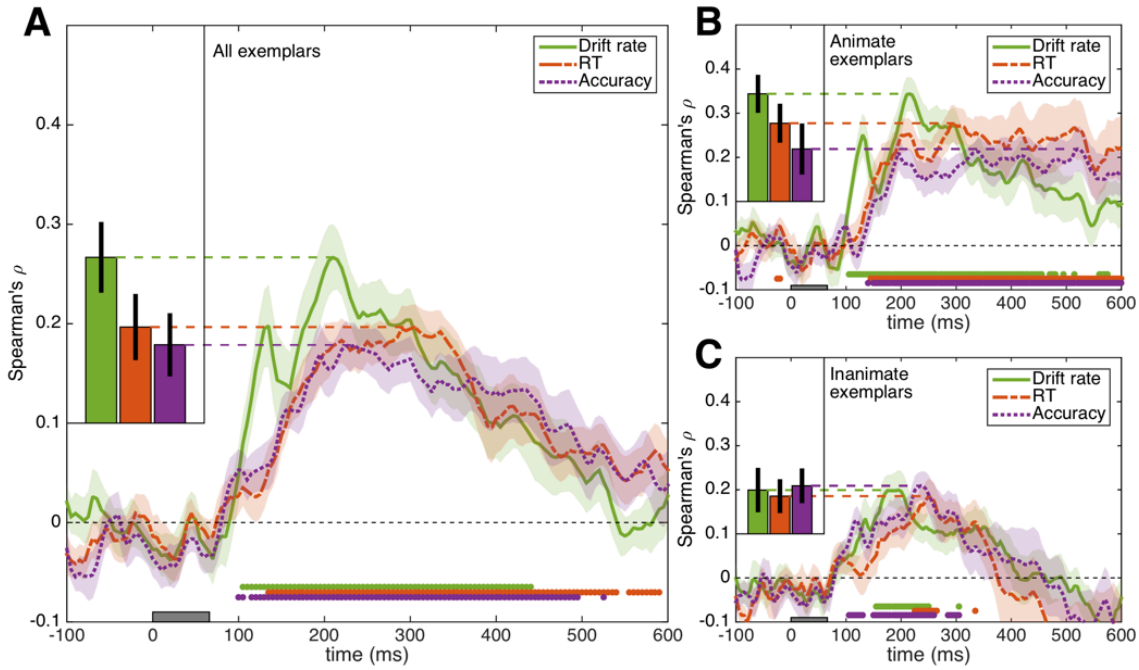

Figure 7. The correlation between distance to boundary, behavioural variables, and drift rate. A. The distance model applied to all the exemplars. **B.** Results for animate exemplars only. **C.** Results for inanimate exemplars. Three measures (RT, accuracy, and the drift rate parameter) were correlated with distance to boundary over time. RT refers to inverted normalized RT. Accuracy is the mean accuracy for exemplars, and drift rate was estimated by fitting LBA. Shaded areas refer to the standard error over subjects, and coloured marks above the x-axis indicate significant above-zero correlations. The peak correlations with their standard errors are compared in the inset (bars top left). The grey bar on the x-axis indicates the time the stimulus was on the screen (0-66ms).

Having established that distance to boundary predicts RT, accuracy and drift rate for animacy categorization, we next investigated the relative contributions of animate and inanimate exemplars. In Ritchie et al., (2015), time-varying correlations for the inanimate exemplars were not significant. Thus the effect was specific to animate exemplars, and the same asymmetry was also reported in Carlson et al., (2014). In this study however, due to the degrading of stimuli, we also predicted a correlation for inanimate exemplars

(when using both clear and degraded exemplars), as we predicted that the representations of degraded inanimate exemplars would still shift towards the classifier decision boundary (Figure 1). Figure 7 shows the result of computing the time-varying correlation separately for animates (Figure 7B) and inanimates (Figure 7C). Animate exemplars reach higher correlations on all behavioural measures, and follow the same trends as seen in Figure 7A for the combined stimulus set. In contrast, inanimate exemplars have lower correlations overall and were less sustained over time. However, significant above-chance correlations between distance and RT, accuracy, and drift rate were still present for the inanimate exemplars. In addition, more sustained correlations (more significant time points) are present for drift rate and accuracy than for RT, for the inanimate exemplars (which is of interest, as previous studies only used RT).

3.4 Comparing the decoding time courses with the time course of predicting behaviour

Ritchie et al. (2015) found that the time-varying correlation with behaviour matched the time-varying classifier decoding performance. To examine whether this relationship was present in the current study, the time-varying correlations with behaviour (RT, accuracy and drift rate) were rank-order correlated to the time-varying decoding result. Note that while a correlation with behaviour explicitly requires above chance decoding (which predicts a correlation between such trajectories), the reverse relationship does not necessarily hold. The results are presented in Figure 8, and show significant correlations between the results from Figure 7 and the decoding trajectory (Figure 6), rising and falling at approximately the same time. When comparing the animate and inanimate trajectories separately (Figure 7B-C), only the animate time-varying correlation matches

decoding performance. These results are consistent with the findings from Ritchie et al. (2015) who also found higher correlations between RT-correlation and decoding trajectories for the animate exemplars. In addition, the decoding trajectory also correlates significantly with the accuracy results from Figure 7, and is significant for both animate and inanimate exemplars. Correlations with the drift rate parameter are also significant, but appear lower in magnitude. This is evident when comparing Figures 6 and 7, where drift rate has a dual peak structure, which differs from the trajectories for the RT and accuracy correlations.

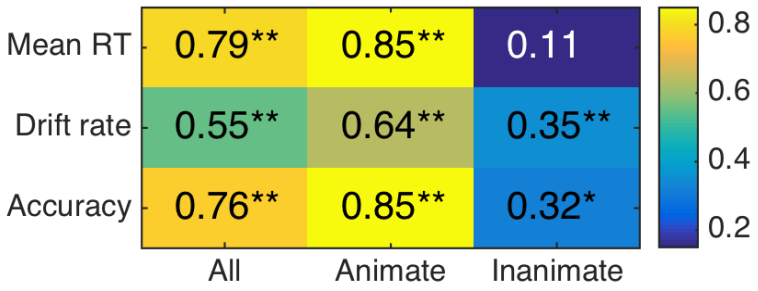

Figure 8. Similarity (Spearman's ρ) between the decoding trajectories (Figure 6), and correlation trajectories (Figure 7). The time varying decoding performance (for all clear and degraded stimuli combined), and time-varying correlations were rank-order correlated. High values indicate similar trajectories (e.g., matching rises, falls, and peaks). Asterisks indicate significant correlations (*=p<.01; **=p<.001).

3.5 Compression of degraded objects in representational space

The RT-distance hypothesis predicts that the neural patterns for degraded objects (which have slower RTs and lower accuracies; see Figure 5) will be closer to the category boundary in representational space, as the distance to the boundary represents the degree of evidence for category membership (Figure 1). The results above show that classifier accuracy is lower for degraded objects, indicating that degraded items are harder to separate in representational space (Figure 6). Next, we showed that distance to boundary correlates with all three behavioural measures (Figure 7A), and this correlation is mostly driven by animate exemplars where it is sustained over time (Figure 7B), with some significant correlations for inanimates across a more limited portion of the time course (Figure 7C). Together, these results seem to favour our prediction based on the RT-distance hypothesis: degraded objects (both animate and, to some extent, inanimate) are closer to the boundary than their clear counterparts, and this shift in representational space correlates with RT, accuracy, and drift rate.

To visually compare our results to the predictions as shown in Figure 1A, we plotted all exemplars in both clear and degraded states relative to the decision boundary in representational space (at the time of peak correlation between drift rate and distance to boundary (210ms), see Figure 7A) at 210ms) in Figure 9. Comparing Figure 1A and Figure 9 suggests that the prediction that degraded stimuli are located closer to the category boundary than clear stimuli holds only for animate, but not inanimate, exemplars. This result is consistent with previous results from Ritchie et al. (2015) and Carlson et al., (2014) which showed no clear relationship between distance to boundary and reaction time for inanimate stimuli. Although, even if the basic relationship between

behaviour and distance to the classifier boundary does not hold for inanimate exemplars, it is still surprising that there is also no relationship between degrading inanimate stimuli and where they are located relative to the category boundary. This asymmetric compression of representational space is however consistent with the interaction between animacy and degrading on the fitted drift rates (Figure 5C), which suggests that there is little or no effect of degrading on the representation of inanimate exemplars.

The compression of exemplar distances towards the animacy boundary is also predicted to match slower drift rates (as shown in Figure 1B). To compare our results to this second prediction, we plotted the mean drift rates over subjects against the mean distance to boundary over subjects for each exemplar in both states in Figure 10 separately for animate (Figure 10A) and inanimate (Figure 10B) exemplars. The result is shown at the time of peak correlation between drift rate and distance to boundary (210ms, Figure 7A). The exemplars are plotted as a function of their mean distance to boundary (x-axis), and mean drift rate (y-axis). The RT-distance hypothesis predicts that exemplars with a higher drift rate are located further away from the boundary. Lines in Figure 10 connect the two states of each exemplar and are coloured green if this prediction is true for each exemplar. This is the case for most of the animate exemplars (Figure 10A), but (not surprisingly, considering the asymmetry in Figure 9) for only a few of the inanimate exemplars (Figure 10B).

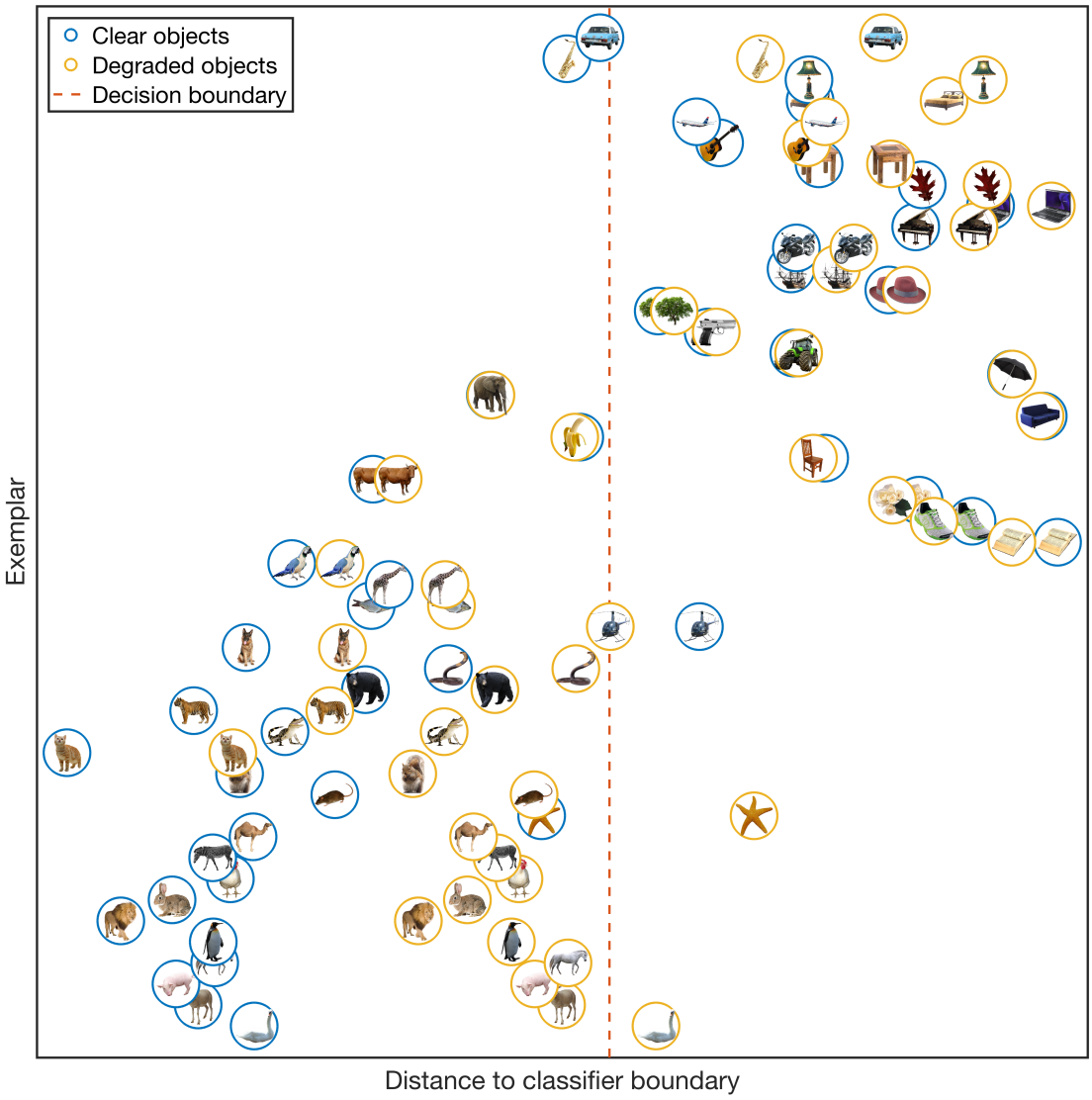

Figure 9. Reconstruction of the representational space for animacy decoding. The location of exemplars in their clear state are plotted in blue circles, and their degraded counterparts in yellow circles. The x-axis represents distance from an object to the boundary, and is the mean distance to the boundary over all subjects at 210ms. The exemplars are ordered on the y-axis according to the shift in distance between their clear and degraded versions, with the largest shift towards the boundary at the bottom of the y-axis, and the largest shift away from the boundary at the top. Note that some objects (e.g., the degraded starfish) have a negative mean distance, and are thus placed on the opposite side of the boundary. Note that degraded animate exemplar representations are in general closer to the linear decision boundary than the clear versions, demonstrating compression of representational space for animate but not inanimate objects.

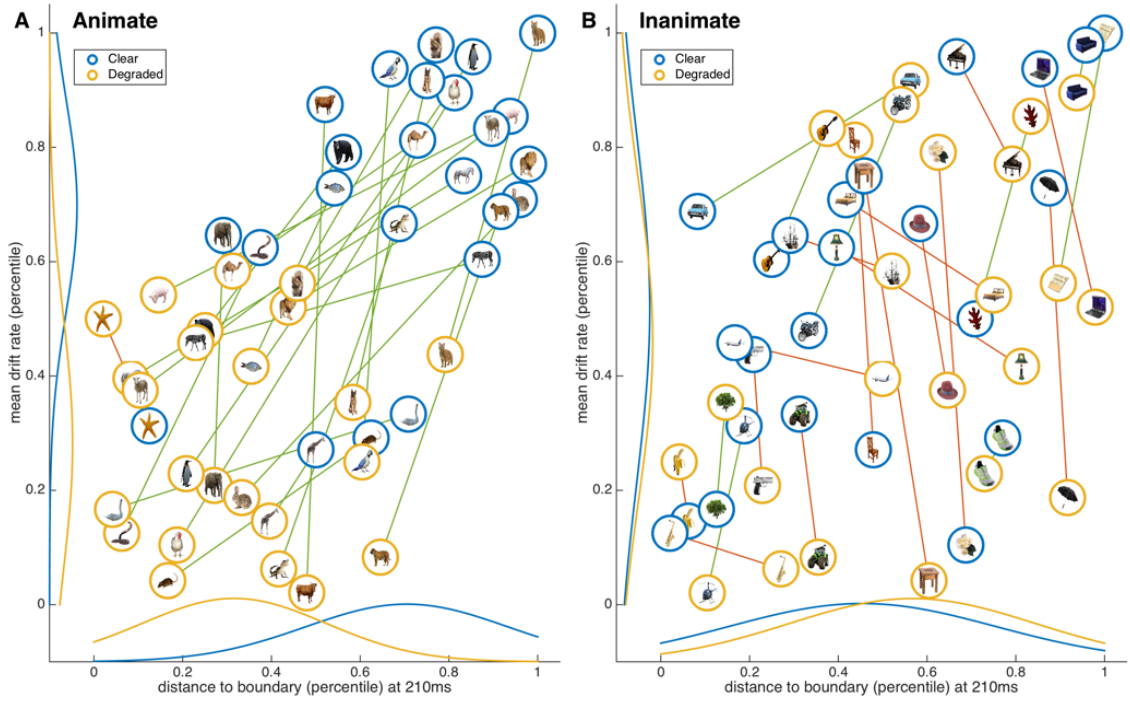

Figure 10. The effect of degrading on the relationship between drift rate and distance to boundary for (A) animate and (B) inanimate exemplars. Items were rank-ordered for mean drift rate over subjects, and mean distance to boundary at the peak LBA prediction time (210ms, see Figure 7A). Here, the y-axis represents the decision boundary. Blue circles are the exemplars in their clear versions, and yellow circles degraded versions. Lines connect the two versions of each exemplar, and are coloured green if the RT-distance hypothesis correctly predicts the direction of the relationship (i.e., the degraded versions of the objects are closer to the boundary and have a slower drift rate), and red if not. The distributions of the exemplars on the variables are shown on the axes.

4 Discussion

Our aim was to test the RT-distance hypothesis in the context of a specific prediction, namely that degrading objects will result in a compression of representational space, and that the shift in representational space for degraded objects will match impaired categorization behaviour. The results showed that degraded object images produced slower RTs, lower accuracies, and slower evidence accumulation rates in an animacy categorization task (Figure 5). When examining the distances of individual objects, we observed an asymmetric compression: only the degraded animate objects had moved closer towards the classifier boundary. Unexpectedly, there was not a corresponding consistent shift towards the boundary for degraded inanimate exemplars. In addition, we found that distance to boundary correlates with drift rate, which is a measure that is more closely related to the decision process than descriptive statistics (mean RT or accuracy).

4.1 Asymmetric effects of stimulus degradation on the neural representation of animacy

We found asymmetrical effects for animate and inanimate objects. Correlations between distance and behaviour (RT/accuracy/drift rate) in this study were driven by animate stimuli. Inanimate stimuli had smaller and less sustained correlations between distance and behaviour, and did not show compression towards the classifier boundary in representational space for degraded versions of the stimuli. Previous studies also found that correlations between distance and RT were almost exclusively driven by animate stimuli (Carlson et al., 2014; Ritchie et al., 2015), which is consistent with our results. Our study differed in many aspects from earlier studies, for example, by using grey scale

stimuli on a controlled background, and excluding human or human face stimuli. Human faces/bodies have faster RTs (Crouzet, Kirchner, & Thorpe, 2010) and highly decodable neural responses (Carlson et al., 2013; Kriegeskorte et al., 2008) and could therefore potentially significantly drive the RT-distance correlations in previous studies. However, we have shown these findings are robust when human faces/bodies are omitted (also note that correlations were calculated within-subject, rather than on the pooled group means as in previous work). Carlson et al. (2014) argue for a conceptual difference between the two animacy categories, suggesting that 'inanimate' is not an equivalent category to animate, for example, because inanimate is negatively defined (i.e., as 'not animate') and is less restricted than the animate category. Furthermore, a clear hierarchical subdivision (e.g., vertebrate - invertebrate) has been reported only within the animate category (Kiani et al., 2007).

We observed that degraded versions of animate objects had compressed towards the boundary, consistent with the RT-distance hypothesis. The lack of compression for the inanimate side of the boundary, and the absence of strong correlations between inanimate categorization behaviour and distance to boundary, suggests that even though the animate/inanimate distinction is highly decodable, it does not sufficiently capture the structure of brain representations linked to responding 'inanimate'. Future work could explore this issue using a different categorization task where both categories are similarly constrained (e.g., faces versus tools), or use a different model of the animacy task (e.g., model it as an animal detection task, instead of animacy categorization). However, it is still possible that responses to other category dichotomies will be based on one category versus 'not' that category. For example a processing bias for faces means they are easier to recognize than tools (cf. Wu, Crouzet, Thorpe, & Fabre-Thorpe, 2015) and an effective

face/tool categorization strategy would be to simply try to detect whether or not a stimulus is a face. In addition, outside the lab, objects are categorized effectively without having to exhaustively test all two-way categorization combinations. Thus, instead of using dichotomous categorization, applying a different task to examine the RT-distance relationship might yield new insights, for example by using go/no-go tasks (Crouzet et al., 2010; Kirchner & Thorpe, 2006; Thorpe, Fize, & Marlot, 1996).

The amount of compression towards the decision boundary was different between individual exemplars (Figure 9). Even though the degraded stimuli were equated for recognizability in an object-naming task, some showed a larger displacement towards the boundary than others (e.g., compare difference between clear and degraded versions of the fish and the sheep in Figure 9). The naming task may be more difficult than the categorization task, and although they likely rely on the same underlying representation (Riesenhuber & Poggio, 2000), different amounts of evidence may be required for naming versus categorization. For example, some animals might be equally easy to name, but when making an animacy decision, some animals are more typically animate than others, which likely makes them easier to categorize than less typical animals. Animacy categorization is known to be influenced by typicality (Posner & Keele, 1968; E. H. Rosch, 1973; E. Rosch & Mervis, 1975). Typical exemplars have also been found to be better decodable (Iordan, Greene, Beck, & Fei-Fei, 2016). For example, mammals such as the cat, tiger, and squirrel are all far from the boundary and have matching fast reaction times and high drift rates. Conversely, the fish and snake are closer to the boundary, and are possibly less typically conceived of as 'animate', for example, because they move and behave differently than mammals. An extreme example is the starfish, which is the animal closest to the boundary in the clear condition,

and is on the wrong side of the boundary (i.e., consistently predicted ‘inanimate’ by the classifier) in the degraded condition, suggesting that it is hard to categorize it as animate (see Figure 9). We found that subjects mostly categorized the starfish as inanimate, and that, when asked, some reported that they do not consider it an animate object. Note that in order to observe a correlation between distance to the boundary and reaction times, distances and reaction times have to systematically differ between exemplars in the same category. Our results thus provide some support for the presence of a more continuous than dichotomous neural representation of animacy in the brain (cf. Connolly et al., 2012; Sha et al., 2015).

Further evidence in support of an animacy asymmetry comes from our observation that inanimate stimuli in general needed to be blurred more to equate their recognizability to animate exemplars (see section 2). This suggests that animate objects are more homogenous than inanimate objects, thus there is greater variance in the 'inanimate' than 'animate’ category (which could be caused by the lack of categorical structure in inanimates). This is consistent with animate objects in general sharing more features than inanimate objects (Garrard, Ralph, Hodges, & Patterson, 2001; McRae, de Sa, & Seidenberg, 1997). The blur filter removes details from the stimuli, but object shape is preserved. It could be the case that the general shape within the broad category of animate objects is more similar, and provides more alternatives (cf. Bracci & op de Beeck, 2016). Even though some higher-level animal subdivisions might be visually dissimilar (e.g., compare fish with birds), subgroups often share similar shapes. For example, outside the laboratory, in order to recognize a blurry zebra (e.g., without glasses), likely alternatives that have similar shapes (e.g., horse, deer, moose) need to be eliminated. In contrast, viewing a blurry piano provides less alternatives which share

the same shape. Together, this supports the notion that inanimate and animate are not equivalent categories, which is consistent with patient studies that found selective deficits in recognition of animate or inanimate objects (Caramazza & Shelton, 1998). Future research could explore whether homogeneity of the category determines the effect of degrading stimuli, by using the distance to bound approach with more homogenous groups of inanimate objects, such as tools or fruits, or restricting the animate category to only domestic animals.

4.2 Drift rate is predicted by distance to boundary

The drift rate parameter from the LBA model directly reflects the speed of evidence accumulation, and builds upon previous work that used RT as a proxy for the decision process (Carlson et al., 2014; Ritchie et al., 2015). We found that correlations with LBA drift rate were on par with those for RT, which was somewhat expected given that RT, accuracy and drift rate are highly correlated. However, there were some qualitative differences in the trajectories. The correlation between LBA drift rate and distance peaked earlier than RT, and had an earlier onset—although this should be interpreted with caution as earlier onsets can be caused by stronger signal-to-noise rather than true underlying differences (Grootswagers et al., 2017). Moreover, drift rate followed a different trajectory than the category decoding trace (as seen in Figure 8), which suggests it may capture a different part of the (neural) decision process. It is interesting that the drift rate trajectory does have a dual peak structure, which resembles previous results of MEG animacy decoding (Carlson et al., 2013; Cichy et al., 2014; Ritchie et al., 2015), but which was not present in our MEG decoding results. This dual-peak structure may suggest that at some point (in the time period between the peaks), distance to

boundary is used to a lesser extent for forming the categorization decision. Alternatively, the decision could have already been made after the first peak, as categorization happens very fast (Crouzet et al., 2010; Kirchner & Thorpe, 2006; Thorpe et al., 1996), and the second peak reflects, for example, a feedback process that has the same representational structure (i.e., the same distances). Taken together, it is sensible to relate distance directly to evidence accumulation using drift rate as it is more closely related to decision processes and therefore it may successfully relate to a wider range of behaviour than RT alone. Future research could explore this further by using tasks that allow more for more variance in RT and accuracy (e.g., incentivizing different speed-accuracy trade-offs), where drift rate is not as highly correlated with RT as in the current study. Moreover, as we showed it is possible to link one decision model with neural distance to boundary, this could be tested with other models of decision making, such as exemplar-based models of choice (Nosofsky & Stanton, 2005) or their LBA-based extension (Donkin & Nosofsky, 2012). More complex models of choice behaviour explicitly parameterize noise fluctuations, and they might better describe the inherently noisy neuroimaging data, but would require more trials per condition to obtain a good fit. Alternatively, different methods of obtaining decision boundaries can be tested. However, linear classifiers perform very similarly on MEG data (Grootswagers et al., 2017) and therefore their decision boundaries will not likely be different. Instead, results from non-linear classifiers can be compared with our results, but they generally return sub-optimal decoding solutions in MEG (Grootswagers et al., 2017).

4.3 The neural dynamics of visual object categorization

We found that significant decoding performance of clear stimuli started around 70ms, and peaked at 345 ms, which was later for degraded stimuli (165 ms and 380 ms, respectively). Note that differences in decoding onsets have to be interpreted with caution, as a later onset can be a result of a lower overall decoding performance (Grootswagers et al., 2017; Isik et al., 2014). Still, we observed that the first local maximum in the decoding trace for the clear objects was absent in the degraded objects (Figure 6). This difference suggests that some information in the early response is predictive of animacy in the clear condition. As the timing of the early peak corresponds to early to mid-level visual areas (Cichy et al., 2016; Thorpe et al., 1996; VanRullen & Thorpe, 2001), the predictive information in the first local maximum in the clear decoding condition could reflect low-level visual information that is removed when degrading the stimuli (Kirchner & Thorpe, 2006).

We found that the peak correlation between distance and drift rate occurred at 210 ms, and the onset of significant correlations with drift rate was at 100 ms. These values did not match the onset or peak of decoding, suggesting that the optimal time for read out does not necessarily correspond with the time that the information can best decoded from the signal. In contrast, Ritchie et al., (2015) found correlations during the whole time period of significant decoding. A possible explanation for this difference is that the fast-paced task and short stimulus duration (66 ms compared to 500 ms in Ritchie et al., (2015)) in the current study promoted a faster read out of animacy by exploiting low level visual cues (Hong, Yamins, Majaj, & DiCarlo, 2016; Kirchner & Thorpe, 2006; Thorpe et al., 1996). This would in addition explain why exemplars such as the banana and

helicopter, which have more rounded shapes than other inanimate objects, are closer to the boundary (Figure 9).

In this study, all MEG channels were included for training and testing the classifier. A limitation of this approach is that the spatial source of the decodable signal is unknown. Because participants were performing an animacy task in the scanner, it is possible that the source of the decoding is the decision process (e.g., frontal executive areas), rather than the representation space in IT. However, as previous research has found no difference in decoding performance between active animacy categorization versus a distractor task (Ritchie et al., 2015). In addition, the representational structure in the MEG response to visual objects has been found to correspond best to fMRI representations in the ventral visual stream, showing that the early and late MEG response respectively matched the V1 and IT fMRI responses (Cichy et al., 2014, 2016). Taken together, these findings show that the most likely sources of information in our study originate from areas in the ventral visual stream, rather than decision making areas.

4.4 Conclusion

In this study we tested whether representational space is compressed when degrading stimuli, and whether this matches behaviour in a dichotomous categorization task. We found that degrading stimuli made them harder to categorize, and that this was accompanied by a compression of representational space, as predicted by the RT-distance hypothesis. This compression was only observed for animate stimuli, suggesting an asymmetry in in the neural representation of animacy. Moreover, we showed that neural distance to boundary can be directly related to a current model of evidence accumulation (LBA) as the fitted drift rates from this model correlated with distance to the boundary. Connecting linear classifiers to models of the decision processes is a step towards relating brain imaging to behaviour, a fundamental and complex challenge in cognitive neuroscience (de-Wit et al., 2016; Forstmann & Wagenmakers, 2015; Purcell & Palmeri, 2016).

Acknowledgements

This research was supported by an Australian Research Council (ARC) Future Fellowship (FT120100816) and ARC Discovery project (DP160101300) awarded to T.A.C. S.G.W. is supported by an Australian NHMRC Early Career Fellowship (APP1072245). The authors declare no competing financial interests.

References

Ashby, F. G. (2000). A Stochastic Version of General Recognition Theory. *Journal of Mathematical Psychology*, *44*(2), 310–329. https://doi.org/10.1006/jmps.1998.1249

Ashby, F. G., & Maddox, W. T. (1994). A Response Time Theory of Separability and Integrality in Speeded Classification. *Journal of Mathematical Psychology*, *38*(4), 423–466. https://doi.org/10.1006/jmps.1994.1032

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64. https://doi.org/10.1016/0010-0285(88)90024-2

Bracci, S., & op de Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. *Journal of Neuroscience*, *36*(2), 432–444. https://doi.org/10.1523/JNEUROSCI.2314-15.2016

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science*, *144*(3617), 424–425.

Caramazza, A., & Shelton, J. R. (1998). Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience*, *10*(1), 1–34. https://doi.org/10.1162/089892998563752

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction Time for Object Categorization Is Predicted by Representational Distance.

Journal of Cognitive Neuroscience, *26*(1), 132–142. https://doi.org/10.1162/jocn_a_00476

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1. https://doi.org/10.1167/13.10.1

Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and Behavior in Decision-Making. *PLOS Comput Biol*, *10*(7), e1003700. https://doi.org/10.1371/journal.pcbi.1003700

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … Haxby, J. V. (2012). The Representation of Biological Classes in the Human Brain. *The Journal of Neuroscience*, *32*(8), 2608–2618. https://doi.org/10.1523/JNEUROSCI.5547-11.2012

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4), 16. https://doi.org/10.1167/10.4.16

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*(5), 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151. https://doi.org/10.1016/j.jmp.2010.10.001

Donkin, C., & Nosofsky, R. M. (2012). A Power-Law Model of Psychological Memory Strength in Short- and Long-Term Recognition. *Psychological Science*, *23*(6), 625–634. https://doi.org/10.1177/0956797611430961

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321. https://doi.org/10.3758/BF03330618

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394. https://doi.org/10.1364/JOSAA.4.002379

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4939-2236-9.pdf

Garrard, P., Ralph, M. A. L., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174. https://doi.org/10.1080/02643290125857

Gaspar, C. M., & Rousselet, G. A. (2009). How do amplitude spectra influence rapid animal detection? *Vision Research*, *49*(24), 3001–3012. https://doi.org/10.1016/j.visres.2009.09.021

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247

Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2016). Typicality sharpens category representations in object-selective cortex. *NeuroImage*, *134*, 170–179. https://doi.org/10.1016/j.neuroimage.2016.04.012

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Joseph, J. E., & Proffitt, D. R. (1996). Semantic versus perceptual influences of color in object recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 407.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309. https://doi.org/10.1152/jn.00024.2007

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776. https://doi.org/10.1016/j.visres.2005.10.002

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press on Demand.

Macé, M. J.-M., Delorme, A., Richard, G., & Fabre-Thorpe, M. (2009). Spotting animals in natural scenes: efficiency of humans and monkeys at very low contrasts. *Animal Cognition*, *13*(3), 405–418. https://doi.org/10.1007/s10071-009-0290-4

Macé, M. J.-M., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *European Journal of Neuroscience*, *21*(7), 2007–2018. https://doi.org/10.1111/j.1460-9568.2005.04029.x

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(3), 608.

Ostergaard, A. L., & Davidoff, J. B. (1985). Some effects of color on naming and recognition of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 579.

Párraga, C. A., Troscianko, T., & Tolhurst, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology*, *10*(1), 35–38. https://doi.org/10.1016/S0960-9822(99)00262-6

Párraga, C. A., Troscianko, T., & Tolhurst, D. J. (2005). The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. *Vision Research*, *45*(25–26), 3145–3168. https://doi.org/10.1016/j.visres.2005.08.006

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural Representation of Task Difficulty and Decision Making during Perceptual Categorization: A Timing Diagram. *The Journal of Neuroscience*, *26*(35), 8965–8975. https://doi.org/10.1523/JNEUROSCI.1655-06.2006

Philiastides, M. G., & Sajda, P. (2006). Temporal Characterization of the Neural Correlates of Perceptual Decision Making in the Human Brain. *Cerebral Cortex*, *16*(4), 509–518. https://doi.org/10.1093/cercor/bhi130

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*(1), 53–68. https://doi.org/10.1037/h0033871

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353.

Purcell, B. A., & Palmeri, T. J. (2016). Relating accumulator model parameters and neural dynamics. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2016.07.001

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226–1243. https://doi.org/10.1037/a0036801

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., & Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *The Journal of Neuroscience*, *33*(18), 7691–7699. https://doi.org/10.1523/JNEUROSCI.3905-12.2013

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212.

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–1204. https://doi.org/10.1038/81479

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience*, 190. https://doi.org/10.3389/fnins.2016.00190

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLoS Comput Biol*, *11*(6), e1004316. https://doi.org/10.1371/journal.pcbi.1004316

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. Retrieved from http://psycnet.apa.org/psycinfo/1974-20611-001

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Sonka, M., Hlavác, V., & Boyle, R. (2008). *Image processing, analysis and and machine vision (3. ed.).* Thomson.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

VanRullen, R., & Thorpe, S. J. (2001). The Time Course of Visual Processing: From Early Perception to Decision-Making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461. https://doi.org/10.1162/08989290152001880

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*(8–9), 1520–1529. https://doi.org/10.1016/j.visres.2005.11.008

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*(6), 685–686. https://doi.org/10.1038/nn1900

Wu, C.-T., Crouzet, S. M., Thorpe, S. J., & Fabre-Thorpe, M. (2015). At 120 msec You Can Spot the Animal but You Don't Yet Know It's a Dog. *Journal of Cognitive Neuroscience*, *27*(1), 141–149. https://doi.org/10.1162/jocn_a_00701

Wurm, L. H., Legge, G. E., Isenberg, L. M., & Luebker, A. (1993). Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 899.

Wyatt, D. F., & Campbell, D. T. (1951). On the liability of stereotype or hypothesis. *The Journal of Abnormal and Social Psychology*, *46*(4), 496–500. https://doi.org/10.1037/h0054690

Chapter 4

Beyond Brain Decoding: Searching for decodable object category information in the brain that predicts behaviour

Tijl Grootswagers[1,2,3], Radoslaw M. Cichy[4], & Thomas A. Carlson[2,3]

[1] Department of Cognitive Science, Macquarie University, Australia

[2] ARC Centre of Excellence in Cognition and its Disorders, Australia

[3] School of Psychology, University of Sydney, Australia

[4] Department of Education and Psychology, Free University Berlin, Germany

Abstract

Using “brain decoding” methods, it has been repeatedly shown that information about a stimulus, such as visual object category, can be decoded from brain activation patterns. An implicit assumption in these studies is that if information can be decoded, then this information is available to the brain for use in behaviour. In the current study, we combined the distance to bound approach analysis to investigate where in the brain decodable visual object category information exists, and additionally, where it is suitable for read out in behaviour, assuming a linear read-out process. We found decodable category information along most of the ventral and dorsal visual streams. However, in only a subset of locations this information could be used to predict observer categorization reaction times, mainly in the anterior Ventral Temporal Cortex (VTC). Our results support the important role the VTC potentially plays in object categorization. Further, they suggested that mid-level ventral and dorsal areas contribute to categorization decisions. By mapping the dissociation between decodable information and information that can be used to predict behaviour, we showed that only a subset of decodable information is relevant for the brain in behaviour.

1 Introduction

Using Multi-Variate Pattern Analysis (MVPA) or "brain decoding" methods, information can be decoded from functional Magnetic Resonance Imaging (fMRI) activation patterns (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Haxby et al., 2001; Haynes, 2015; Kamitani & Tong, 2005; Kriegeskorte, Goebel, & Bandettini, 2006; Tong & Pratte, 2012). Many MVPA studies have been conducted in the domain of visual object recognition, showing that categorical information, such as object animacy, can be reliably decoded from fMRI Ventral Temporal Cortex (VTC) patterns of activation (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014, 2016; Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008; O'Toole, Jiang, Abdi, & Haxby, 2005; Sha et al., 2015). An implicit assumption in many brain decoding studies is that if information can be decoded, then this information is available to the brain to use in behaviour (de-Wit, Alexander, Ekroll, & Wagemans, 2016; Ritchie, Kaplan, & Klein, in press). However, this does not have to be the case, as it could be that the decoded information is merely a by-product of a different signal that is relevant for the brain (de-Wit et al., 2016; Williams, Dang, & Kanwisher, 2007).

A fundamental challenge in cognitive neuroscience is to relate decoded information represented in the brain to behaviour (de-Wit et al., 2016). To date, studies have approached this issue by correlating classifier performances to behaviour (Naselaris, Kay, Nishimoto, & Gallant, 2011; Raizada, Tsao, Liu, & Kuhl, 2010; van Bergen, Ji Ma, Pratte, & Jehee, 2015; Walther, Caddigan, Fei-Fei, & Beck, 2009; Williams et al., 2007), or by comparing the similarity structure of the decoded information to behavioural similarity ratings (Bracci & op de Beeck, 2016; Cohen, Dennett, &

Kanwisher, 2016; Mur et al., 2013; Proklova, Kaiser, & Peelen, 2016; Redcay & Carlson, 2015; Wardle, Kriegeskorte, Grootswagers, Khaligh-Razavi, & Carlson, 2016). Another approach to address this issue rests on the similarity between the MVPA classifiers and evidence accumulation models of human decision making under signal detection theory (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Ritchie & Carlson, 2016; Ritchie, Tovar, & Carlson, 2015). Classifiers in decoding studies fit multi-dimensional hyperplanes to separate the neural activation space into two classes and use this hyperplane to predict the class of new items. If the brain in a decision-making task uses a linear read-out process (DiCarlo & Cox, 2007) using the same neural activation space, then this decision hyperplane would reflect an observer's decision boundary for that decision. In classic signal detection theory, the closer the input is to this decision boundary, the more ambiguous the evidence for the decision (Green & Swets, 1966). Thus, if for a decision task (e.g., categorization), the brain 'uses' the same information as the MVPA classifier, then the distance to the classifier hyperplane reflects evidence for the decision task. Ambiguous evidence in turn predicts longer reaction times (Ashby, 2000; Ashby & Maddox, 1994), which predicts that distance to the classifier hyperplane negatively correlates with reaction time.

Here, we examined the relationship between decodable information and behaviour in a two-step approach. First, we used MVPA classifiers to show where decodable category information exists, using the searchlight method on fMRI data (Haynes et al., 2007; Kriegeskorte et al., 2006). Secondly, we asked where the distance of the individual stimuli to the classifier's hyperplane predicts reaction times for those stimuli in the same categorization task. This approach yielded decoding and correlation maps of areas involved in visual object categorization. We applied this analysis to two fMRI

datasets from independent experiments (Cichy et al., 2014, 2016) that had different stimulus sets, which allowed for testing the robustness of the results. We tested several different categorization contrasts, to assess the generalizability of the neural distance to bound approach. Our results show that decodable information exists along the entire ventral and dorsal visual streams, but that behaviour can only be predicted from a subset of those locations, mainly in the ventral-occipital-temporal cortex. These results clearly indicate a distinction between decodable information, and information that could be used by the brain for behaviour.

2 Methods

The aim of this study was to correlate distance to a classifier's hyperplane to reaction times (RTs) for categorization. In this section, we first describe the stimuli and fMRI data that were obtained from previous experiments (Cichy et al., 2014, 2016). Next, we collected categorization reaction times for those stimuli on Amazon's Mechanical Turk. Finally, we describe the two-step searchlight procedure used to create decoding and correlation maps of areas involved in visual object categorization.

2.1 Stimuli

We used the stimuli from two experiments that have been published (Cichy et al., 2014, 2016). Stimuli for experiment 1 consisted of 92 visual objects, segmented on a white background (Figure 1A). Stimuli consisted of animate and inanimate objects. The animate objects could be further divided into faces, bodies, humans and animals. Inanimate objects consisted of natural (e.g., plants or fruits) and man-made items (e.g.,

tools or houses). The stimulus set for experiment 2 consisted of 118 visual objects on natural backgrounds, shown in Figure 1C. A small proportion of the objects (27) were animate. The inanimate objects included subcategories such as tools, or food items. In both experiments, participants were presented with the visual object stimuli while performing an orthogonal task at fixation. Stimuli were displayed at 2.9° (Experiment 1) and 4.0° (Experiment 2) visual angle with 500 ms duration. Images were displayed (overlaid with a grey fixation cross) for 500 ms in random order. Participants performed an orthogonal task while in the scanner: on 25% of the trials no image was displayed, and instead, participants responded with a button press to the fixation cross turning darker.

2.2 fMRI recordings

The first experiment (for a complete description, see Cichy et al., 2014) had high resolution fMRI coverage of the ventral visual stream (Figure 1B) from 16 participants with a 2 mm isotropic voxel resolution. The second experiment (for a complete description, see Cichy et al., 2016) had whole brain recordings from 15 participants with a 3 mm isotropic voxel resolution. In both experiments, at the start of a session, structural images were obtained using a standard T_1-weighted sequence. fMRI data were aligned and coregistered to the T1 structural image, and then normalized to a standard MNI template. The data were not smoothed. A general linear models (GLM) was used to estimate responses for each stimulus (92 and 118, respectively). Movement parameters were included in the GLM as nuisance parameters, and stimulus onset and duration were used as regressors and convolved with a

hemodynamic response function. The estimated GLM parameter for each stimulus were then contrasted against baseline to obtain t-values for each stimulus.

2.3 Reaction time data

We obtained reaction times for the stimuli in multiple different categorization contrasts (see Figure 1A&B). For the 92-data, these were animate - inanimate, face - body, and human - animal. For the 118-data, we tested animate - inanimate, tool - other, food - other, and transport - other. The RTs were collected using Amazons Mechanical Turk (MTurk). For each of the categorization contrasts, 50 unique participants performed a categorization task using the same stimuli as were used in collecting the fMRI data. Participants were instructed to "Categorize the images as fast and accurate as possible using the following keys: (z for X, m for Y)", where X and Y would be replaced with the relevant categories (e.g., animate and inanimate) for the contrast. On each trial, an image was presented for 500ms, followed by a black screen until the participant's response (Figure 1C). The presentation order of the stimuli was randomized and stimuli did not repeat. This resulted in 50 reaction time values per exemplar (one for each participant). Each participant's reaction times were z-scored. Next, we computed the median reaction time (across participants) for each exemplar. This resulted in one reaction time value per exemplar, which were used in the rest of the study.

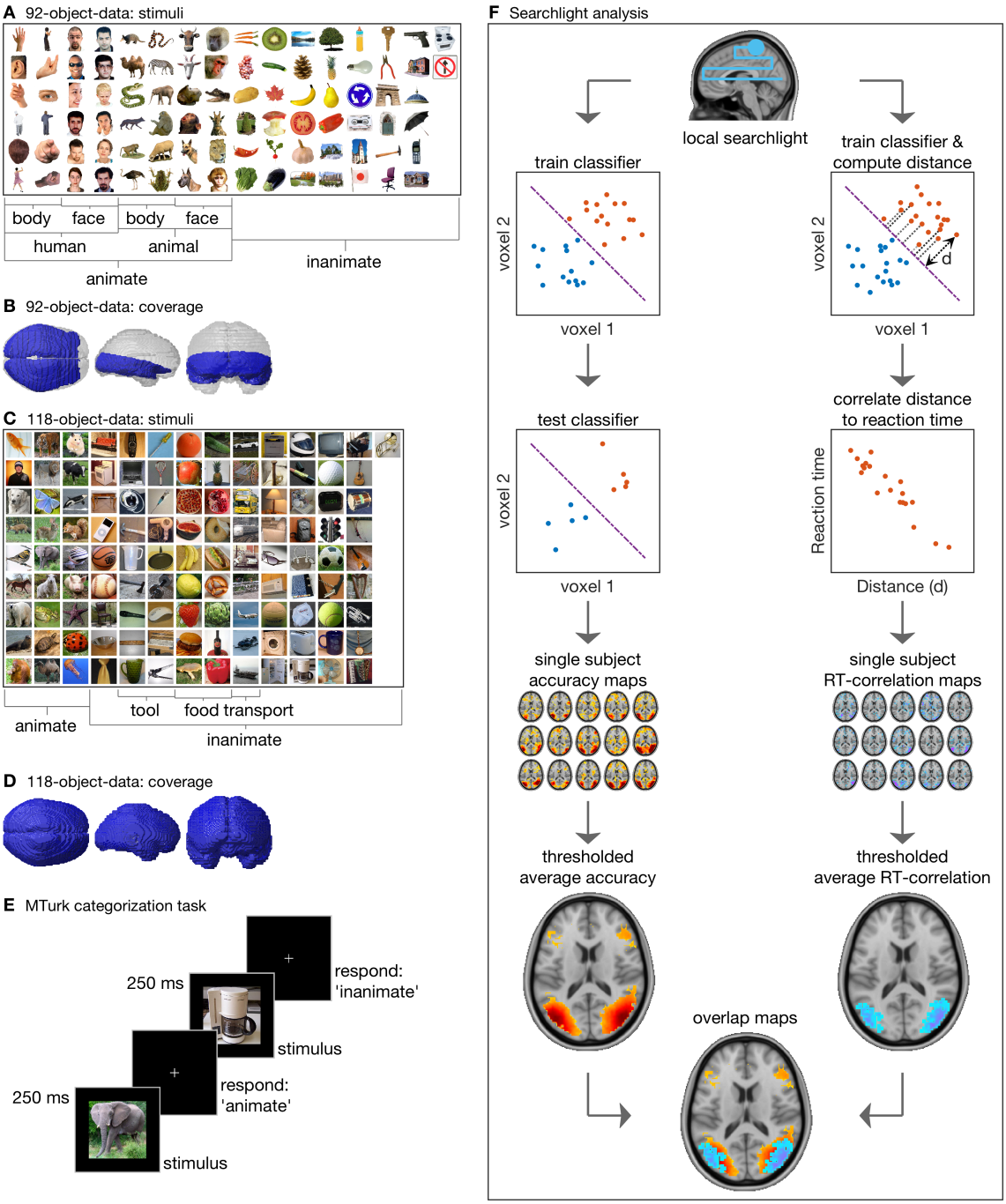
A 92-object-data: stimuli
body
face
body
face
human
animal
animate
inanimate
B 92-object-data: coverage
C 118-object-data: stimuli
animate
tool
food transport
inanimate
D 118-object-data: coverage
E MTurk categorization task
250 ms
respond:
'inanimate'
stimulus
250 ms
respond:
'animate'
stimulus
F Searchlight analysis
local searchlight
train classifier
voxel 2
voxel 1
test classifier
voxel 2
voxel 1
single subject
accuracy maps
thresholded
average accuracy
train classifier &
compute distance
voxel 2
d
voxel 1
correlate distance
to reaction time
Reaction time
Distance (d)
single subject
RT-correlation maps
thresholded
average RT-correlation
overlap maps

Figure 1. Illustration of the searchlight approach used in this study including stimuli, task, and brain coverage for both datasets. A, B). Stimuli (**A**) and coverage (**B**) in fMRI study 1. **C, D)** Stimuli (**C**) and coverage (**D**) in fMRI study 2. **E.** Independent reaction times for categorization contrasts were collected on MTurk. On each trial, a stimulus was displayed for 250ms, and participants categorized it into two categories (here: animate vs inanimate) with a keypress. **F.** The searchlight analysis procedure consisted of two separate analyses to create accuracy maps, and distance-RT-correlation maps. For each voxel, a local cluster of neighbouring voxels was used to train and test a classifier to decode the contrast (e.g., animacy), and its accuracy was stored at the centre voxel of the local cluster. At the same local voxel cluster, a classifier was trained to obtain a decision hyperplane. The distances of each stimulus to this hyperplane were correlated to the independent reaction times obtained on MTurk, and this value was stored at the centre voxel. Repeating the procedure over all voxels resulted in an accuracy and correlation map. The procedure was repeated for all subjects and the results were averaged at the group level and thresholded at $p_{corrected} < 0.05$. Finally, for visualization, the correlation maps were superimposed on the accuracy maps.

2.4 Searchlight procedure

For each categorization contrast and subject, we used a searchlight approach (Haynes et al., 2007; Kriegeskorte et al., 2006) to create maps of decoding accuracy and of correlations between distance to the classifier boundary and categorization reaction times. In contrast to pre-defined ROI's, which are used to test a-priori hypotheses about the spatial origin of information in the brain, the searchlight results in a spatially unbiased map of decodable information. An overview of the approach is presented in Figure 1D.

To create the decoding accuracy maps, we used a standard searchlight decoding approach (Kriegeskorte et al., 2006), as implemented in the CoSMoMVPA decoding

toolbox (Oosterhof, Connolly, & Haxby, 2016). For each voxel in the three-dimensional fMRI voxel space, t-values belonging to a spherical cluster (4 voxel radius) centred around this voxel were arranged into a vector for each stimulus. The stimuli vectors were then split up into five sets for cross-validation. A support vector machine (SVM) classifier was trained on four sets and its prediction accuracy was tested on the left out set. This was repeated five times, using each set for testing once. The mean accuracy across these repetitions was stored at the centre voxel of the sphere. Repeating the process over all voxels resulted in a 3D-map of decoding accuracies. The searchlight procedure was conducted for each subject independently, resulting in one accuracy map for each subject for each task.

Next, we created maps of correlations between distance to the classifier hyperplane and observer categorization behaviour on the same contrast. Vectors of t-values were obtained for each stimulus in the same manner as the first searchlight procedure. Then, an SVM classifier was trained using all the stimuli, effectively fitting a hyperplane that distinguished between vector patterns according to category. We then computed the distance of each exemplar to this decision hyperplane, resulting in one distance value per exemplar. The distance values were rank-order correlated (Spearman's ρ) to the median MTurk reaction times collected in the behavioural experiment for the same exemplars in the analogous categorization task. To assess the contribution of each category independently, the correlations were performed separately for the two sides of the categorization (i.e., one correlation for animate and one for inanimate exemplars). In an animacy contrast, exemplars from the animate category have been previously found to contribute highest to the correlations (Carlson et al., 2014; Ritchie et al., 2015). For each categorization task this resulted in two correlation maps per subject.

2.5 Statistical analysis

The resulting maps of decoding accuracy and correlations were assessed for significance at the group level using sign-permutation tests for random-effects inference (Maris & Oostenveld, 2007), which were corrected for multiple comparisons at the cluster level using threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009). First, a permutation distribution was obtained by randomly swapping the sign of the decoding-minus-chance or correlation results 10,000 times. The maximal TFCE-statistic (over all voxels) of these permuted results formed the null-distribution. P values were then computed by comparing the observed TFCE values to the null-distribution, and thresholded at $p_{corrected} < 0.05$.

2.6 Relating the results to topographical locations in the visual system

For each of the categorization contrasts, we identified the locations of the significant clusters with respect to ROIs of the visual system. The significant clusters in the decoding maps and correlation maps were compared to probabilistic topographic maps of visual processing areas (Wang, Mruczek, Arcaro, & Kastner, 2015), which represent for each voxel the visual area with the highest probability. We computed the overlap between the clusters and the ROIs as the number of voxels in each visual ROI that were part of the significant cluster. This allows quantifying the difference in size between decoding clusters and correlation clusters per visual ROI.

2.7 Visualizing the contribution of individual exemplars

For each contrast, we examined the representational space of objects at the local voxel sphere that had the highest correlation between distance and RT. To avoid circularity, we used a leave-one-subject-out approach to obtain the average distances; for each subject, we selected the voxel that had the highest average correlation in the remaining subjects. Next, the distances at this voxel for the left out subject were extracted. We repeated this over all subjects, and averaged the resulting distances over subjects, resulting in one distance estimate per exemplar. We then scattered the stimuli according to their distance-percentiles and their median RT-percentiles. This allowed us to explore how certain types of stimulus (e.g., mammals, faces, or invertebrates) are contributing to the correlations.

3 Results

We created unbiased spatial maps of areas that showed decodable visual object category information and of areas where this decodable information correlates with reaction time in a categorization task. We first performed a decoding searchlight, which yielded a map of significant decoding accuracy in the brain. Then, in another searchlight, we applied the neural distance to bound approach to the individual subjects, correlating the distance to the classifier's hyperplane for each subject to a set of independent RTs for the same categorization task. We will present the results by categorization contrast, starting with animacy.

3.1 Object animacy

We applied the two searchlight analyses to show where in the brain decodable animacy exists and where the decoded information can be used to predict animacy categorization behaviour. Object animacy has been shown before to be a robust organizing principle in VTC representations (Caramazza & Shelton, 1998; Grill-Spector & Weiner, 2014; Kriegeskorte et al., 2008). The first searchlight showed voxel clusters with decodable information along the entire ventral visual stream from the occipital pole to anterior ventral temporal cortex (Figure 2). With the second searchlight, we found that for a subset of these voxels, the distance to the classifier boundary correlated with categorization reaction times for the animate exemplars (Figure 2). Cluster extent and peak location are shown in Table 1A. No significant correlations were found for the inanimate exemplars. Next, we tested the generalizability of these results with a new stimulus set (the 118-object-data), where in addition, we examined both the ventral and dorsal visual streams. For this we used the same approach on the 118-object data which had full brain coverage. As was the case for the 92-object data, a subset of the voxels with decodable information had a significant RT-correlation. In addition, significant RT-correlations were found in the dorsal stream (Figure 2). Even though the stimuli for the 118-object data consisted largely of inanimate objects, no significant correlations were present when performing the correlation using only the inanimates. Cluster extent and peak location are shown in Table 2A. This is consistent with previous findings (Carlson et al., 2014; Ritchie et al., 2015), who also reported correlations mainly driven by animates. Our results build upon these results by showing that the effects replicate with a different stimulus set (the 118-object-data). In sum, we found that only in a subset of the areas with decodable information, this information

correlates to behaviour. This was consistent across stimulus sets. These results suggest that not all decodable information is equally relevant for behaviour.

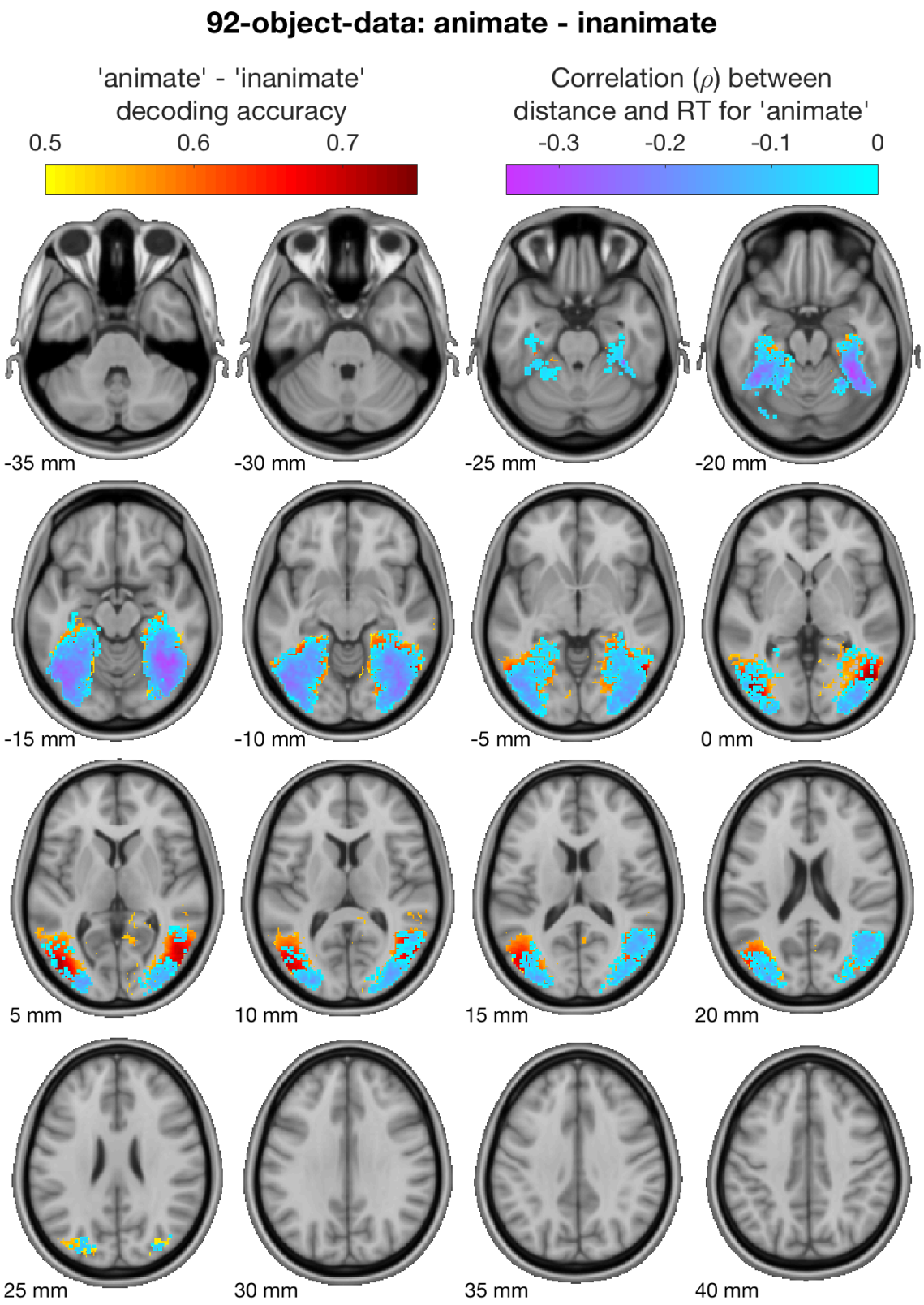

Figure 2. Searchlight results for animacy categorization on the 92-object-data. Clusters with significant decoding are represented in hot colours. Clusters where (in addition to decoding) distance to the classifier's hyperplane correlates significantly with observer reaction times are shown in cool colours. Individual subject results (N=15) were averaged and thresholded at $p_{corrected}<0.05$. The decoding and correlation results are projected onto axial slices standard T_1 image in MNI space. The results show that animacy decoding extends far into the ventral stream, and correlations with behaviour are only present for a subset of these areas.

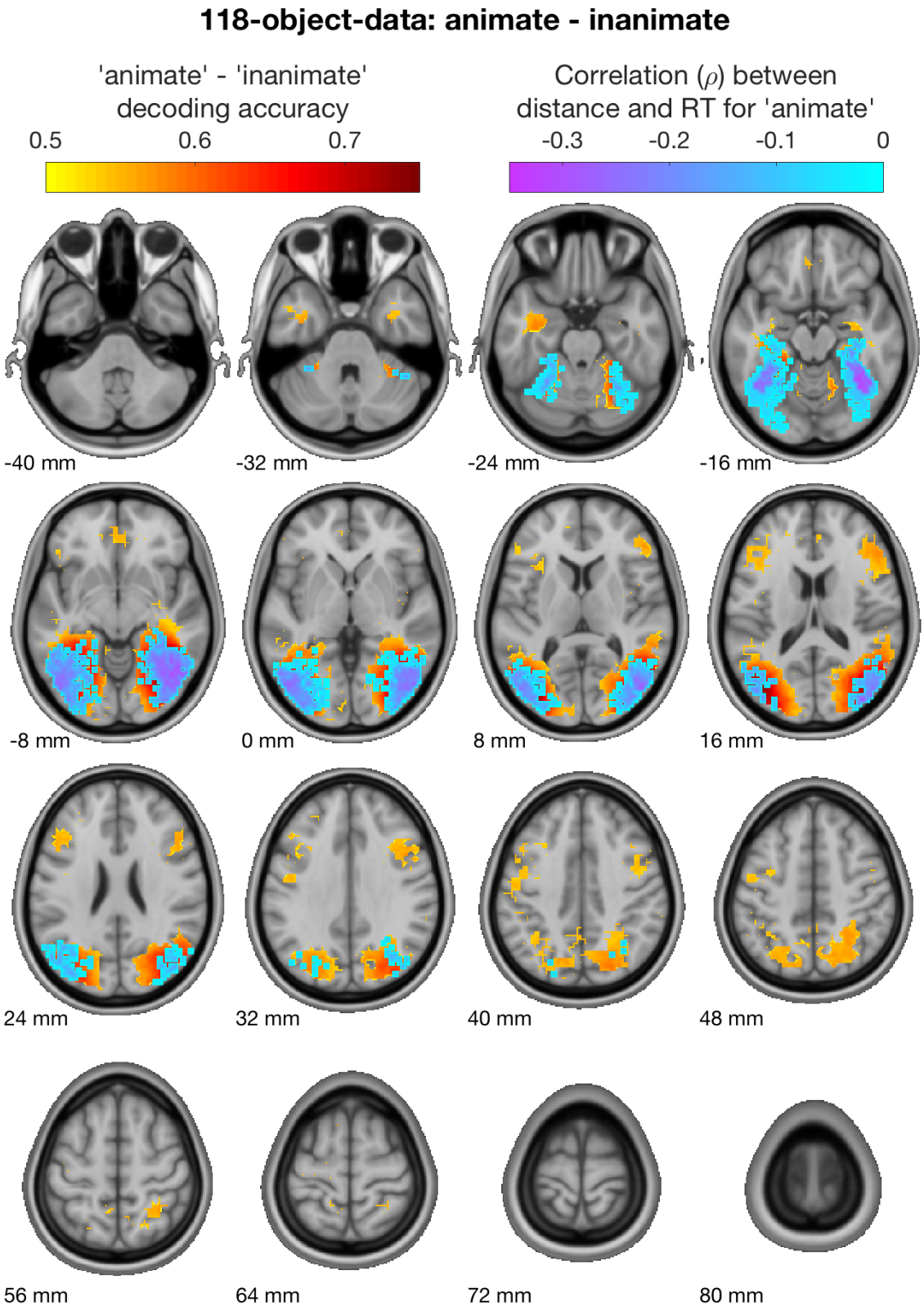
118-object-data: animate - inanimate
'animate' - 'inanimate'
decoding accuracy
0.5
0.6
0.7
Correlation (ρ) between
distance and RT for 'animate'
-0.3
-0.2
-0.1
0
-40 mm
-32 mm
-24 mm
-16 mm
-8 mm
0 mm
8 mm
16 mm
24 mm
32 mm
40 mm
48 mm
56 mm
64 mm
72 mm
80 mm

Figure 3. Searchlight results for animacy categorization on the 118-object-data. Clusters with significant decoding are represented in hot colours. Clusters where (in addition to decoding) distance to the classifier's hyperplane correlates significantly with observer reaction times are shown in cool colours. Individual subject results (N=15) were averaged and thresholded at $p_{corrected}<0.05$. The decoding and correlation results are projected onto axial slices standard T_1 image in MNI space. These results replicate the results for the 92-object data, showing large clusters containing decodable information in the ventral stream, with only a subset of those clusters exhibiting significant correlations with RT. In addition, we observed clusters with decodable information in the prefrontal areas and in the dorsal stream, and correlations with behaviour in the dorsal stream.

3.2 Alternative categorization tasks

We next investigated categorization tasks at a lower tier than animacy. Previous studies have shown areas with preferential responses to for example faces (Kanwisher, McDermott, & Chun, 1997) and bodies (Downing, Jiang, Shuman, & Kanwisher, 2001; Downing & Peelen, 2016). To date, the distance to bound approach has only been applied on the top level animacy distinction (Carlson et al., 2014; Ritchie et al., 2015). Animacy is a large and robust effect (Carlson et al., 2013; Downing, Chan, Peelen, Dodds, & Kanwisher, 2006; Grill-Spector & Weiner, 2014; Kriegeskorte et al., 2008), and it is therefore important to show that the approach generalizes to categorization at the subordinate level, which could be using information in smaller areas in the brain (Downing et al., 2001; Downing & Peelen, 2016; Kanwisher et al., 1997).

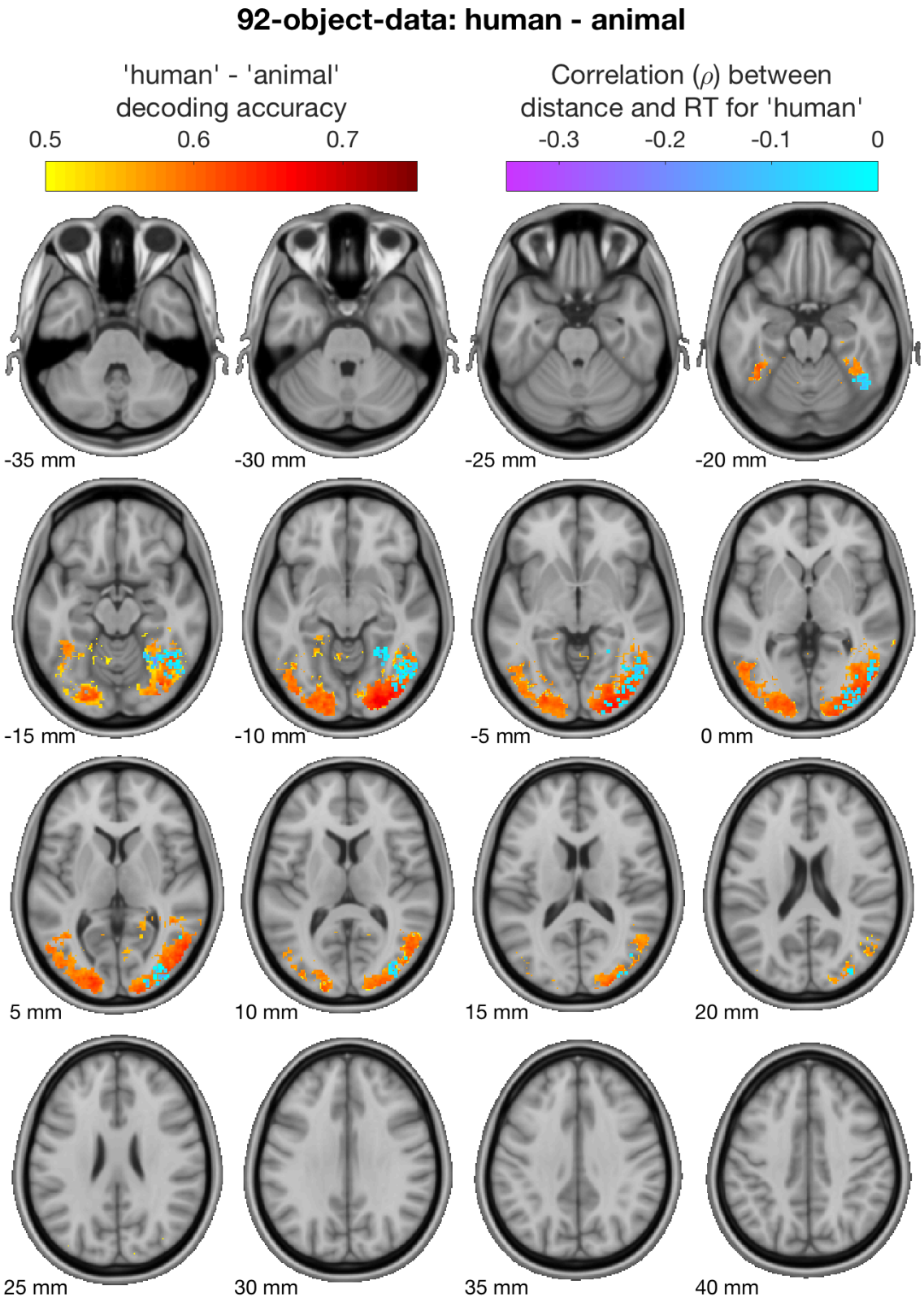
92-object-data: human - animal
'human' - 'animal'
decoding accuracy
0.5
0.6
0.7
Correlation (ρ) between
distance and RT for 'human'
-0.3
-0.2
-0.1
0
-35 mm
-30 mm
-25 mm
-20 mm
-15 mm
-10 mm
-5 mm
0 mm
5 mm
10 mm
15 mm
20 mm
25 mm
30 mm
35 mm
40 mm

Figure 4. Searchlight results for the human-animal categorization contrasts on the 92-object-data. Clusters with significant decoding are represented in hot colours. Clusters where (in addition to decoding) distance to the classifier's hyperplane correlates significantly with observer reaction times are shown in cool colours. Individual subject results (N=15) were averaged and thresholded at $p_{corrected}<0.05$. The decoding and correlation results are projected onto axial slices standard T_1 image in MNI space.

We tested two alternative contrasts on the 92-object data: human – animal (Figure 4), and face – body (Figure 5), to investigate categorization read-out at the subordinate level. We found that a small subset of decodable information had a significant RT-correlations for one category in both tasks, faces and humans respectively. These clusters were much smaller than the decoding clusters, and much smaller than those for the animacy contrast (Table 1). For the alternative tasks (food, transport or tool versus 'other') in the 118-data, we did not find a correlation for the target (food, transport or tool) categories, but we found correlations for the 'other' category in food – other and tool –other, but not in transport – other (Table 2 B-D). Taken together, these results show that the distance to bound approach can be used in categorization contrasts at other levels than animacy.

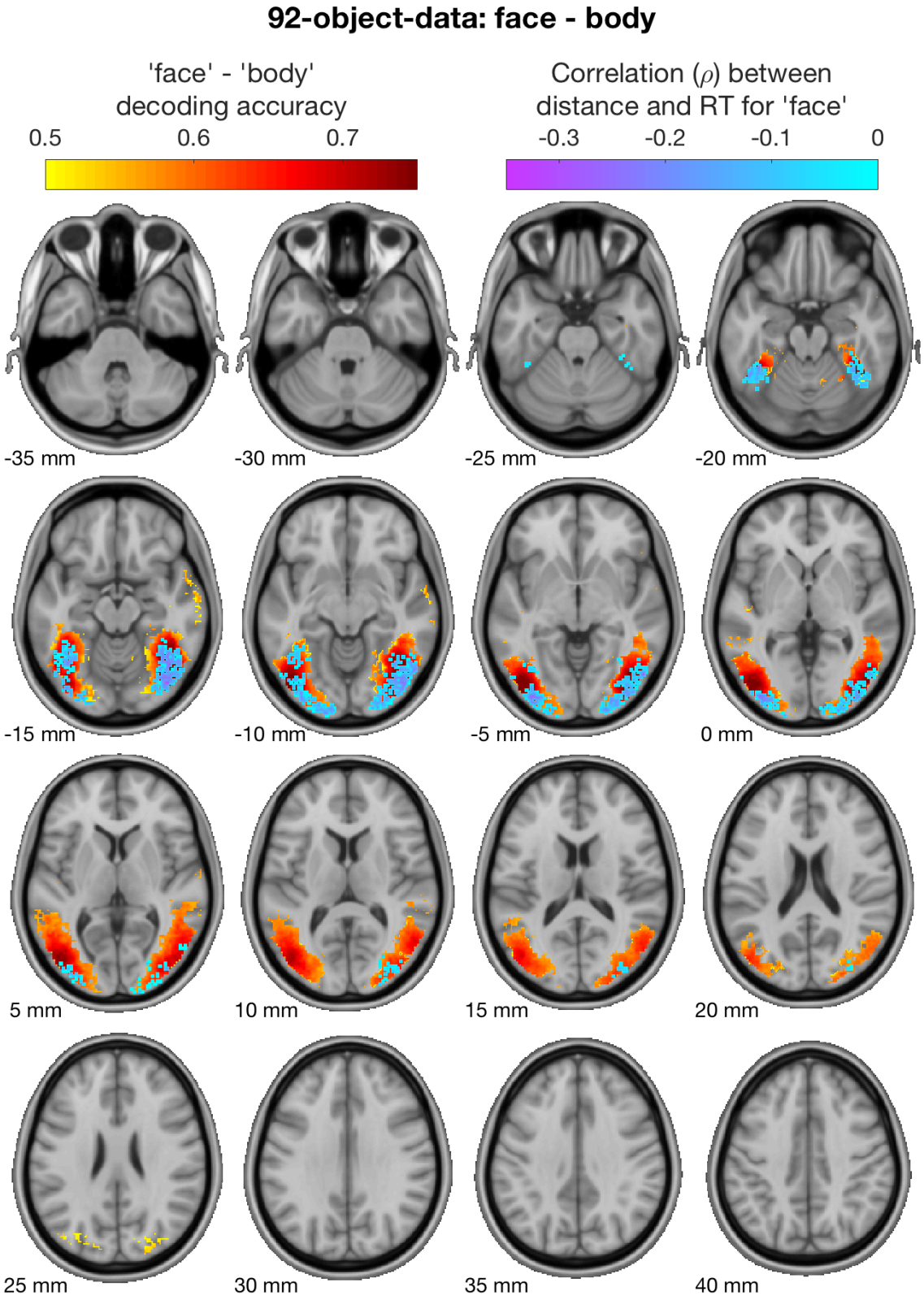
92-object-data: face - body
'face' - 'body'
decoding accuracy
0.5
0.6
0.7
Correlation (ρ) between
distance and RT for 'face'
-0.3
-0.2
-0.1
0
-35 mm
-30 mm
-25 mm
-20 mm
-15 mm
-10 mm
-5 mm
0 mm
5 mm
10 mm
15 mm
20 mm
25 mm
30 mm
35 mm
40 mm

Figure 5. Searchlight results for the face-body categorization contrasts on the 92-object-data. Clusters with significant decoding are represented in hot colours. Clusters where (in addition to decoding) distance to the classifier's hyperplane correlates significantly with observer reaction times are shown in cool colours. Individual subject results (N=15) were averaged and thresholded at $p_{corrected}<0.05$. The decoding and correlation results are projected onto axial slices standard T_1 image in MNI space.

Contrast	**Cluster**	**Size**	**Max/Min**	**TFCE**	**X**	**Y**	**Z**
A) *decoding 'animate' vs 'inanimate'*	1	6997	0.80	3.7190	36	-52	-15
	2	6311	0.77	3.7190	-44	-54	-19
	3	36	0.56	1.7291	-10	-76	-4
RT-correlation 'animate'	1	3515	-0.38	-3.7190	38	-58	-19
	2	2979	-0.32	-3.7190	-36	-66	-17
B) *decoding 'human' vs 'animal'*	1	4619	0.69	3.7190	22	-90	-13
	2	3350	0.65	3.5401	-16	-94	7
	3	22	0.60	1.7636	66	-46	-2
	4	17	0.57	1.6912	-12	-58	7
RT-correlation 'human'	1	208	-0.29	-2.1890	30	-58	-15
	2	72	-0.21	-1.7903	22	-94	-2
C) *decoding 'face' vs 'body'*	1	6089	0.84	3.7190	44	-78	-10
	2	5633	0.79	3.7190	-42	-76	-10
	3	60	0.61	1.9095	-46	-20	-2
	4	36	0.57	1.7660	50	28	-2
RT-correlation 'face'	1	795	-0.32	-3.1947	40	-76	-15
	2	308	-0.28	-2.5302	-36	-86	-6
	3	176	-0.25	-2.3824	-36	-60	-17

Table 1. Cluster details for all contrasts on the 92-data. For all contrasts, all clusters larger than 10 voxels are listed. For each cluster, we report its size in voxels, its peak value (maximum for decoding or minimum for distance-RT correlation), the TFCE statistic for the peak, and the peak's location in MNI-XYZ coordinates. Animacy (**A**) showed the largest clusters and highest decoding accuracy and RT-correlations. The alternative contrasts (**B,C**) resulted in similar sized decoding clusters, but much smaller clusters for the RT-correlations.

Contrast	**Cluster**	**Size**	**Max/Min**	**TFCE**	**X**	**Y**	**Z**
A) *decoding 'animate' vs 'inanimate'*	1	10056	0.80	3.7190	36	-55	-11
	2	735	0.59	2.2904	45	11	31
	3	20	0.54	1.7495	-27	-22	64
RT-correlation 'animate'	1	1133	-0.30	-3.5401	-42	-49	-14
	2	1110	-0.34	-3.7190	51	-73	-2
	3	68	-0.27	-2.2203	36	-55	-26
B) *decoding 'tool' vs 'other'*	1	325	0.58	2.7589	-30	-94	7
RT-correlation 'other'	1	309	-0.20	-3.4316	-39	-85	4
C) *decoding 'transport' vs 'other'*	1	51	0.58	1.8793	-18	-97	-5
D) *decoding 'food' vs 'other'*	1	908	0.61	3.1214	-33	-55	-17
	2	330	0.62	2.6045	36	-55	-14
	3	47	0.57	2.1084	27	-70	31
RT-correlation 'other'	1	294	-0.12	-2.4783	-27	-85	-2
	2	23	-0.13	-2.1890	27	-40	-14

Table 2. Cluster details for all contrasts on the 118-data. For all contrasts, all clusters larger than 10 voxels are listed. For each cluster, we report its size in voxels, its peak value (maximum for decoding or minimum for distance-RT correlation), the TFCE statistic for the peak, and the peak's location in MNI-XYZ coordinates. Animacy (**A**) showed the largest clusters and highest decoding accuracy and RT-correlations, replicating the result on the 92-data. The alternative contrasts (**B-D**) did not show significant correlations for the target category, but only for the 'other' category

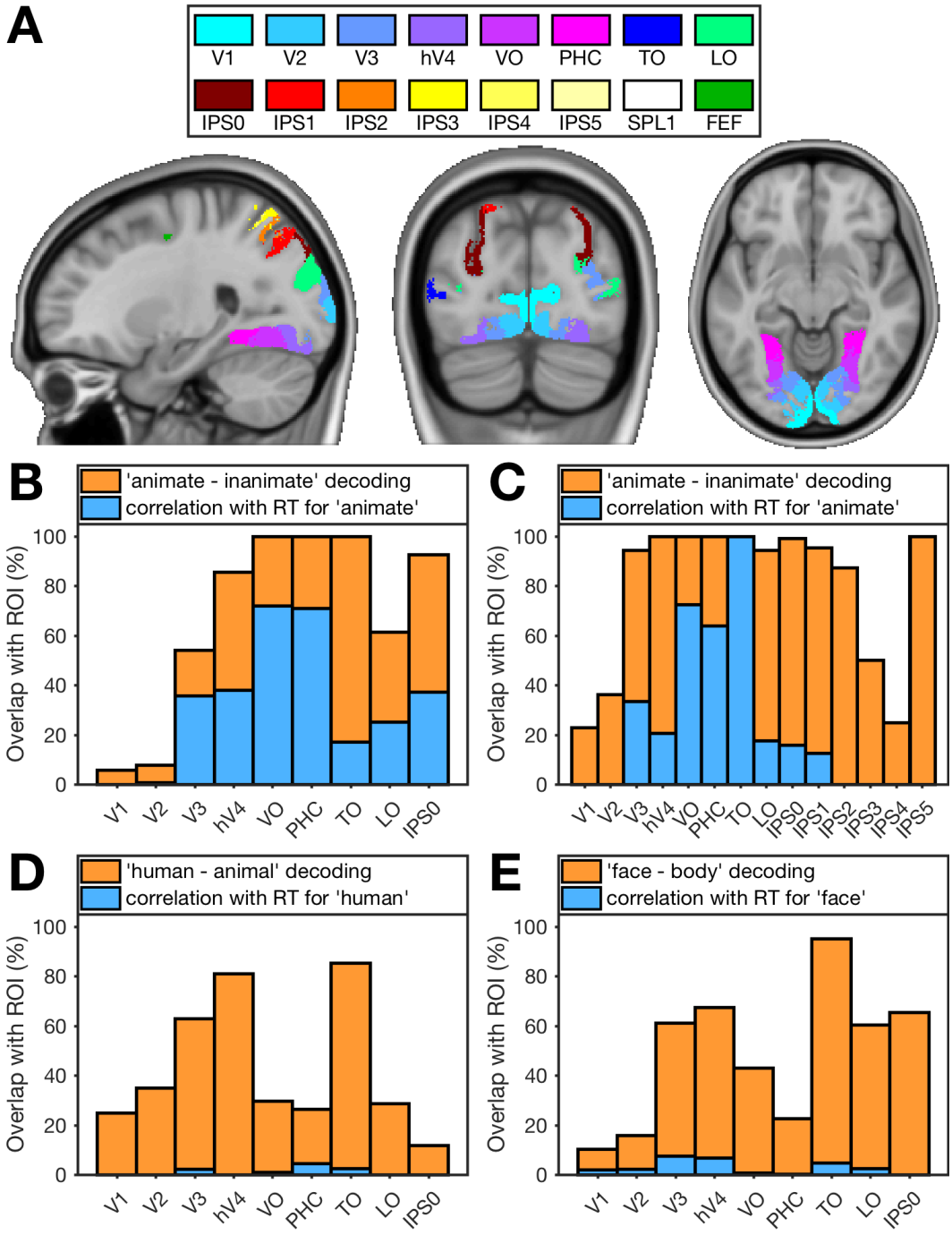

Figure 6. Overlap between our results and visual ROIs. A. Locations of topographical ROIs of the visual system. **B-D.** The percentage overlap between significant clusters and the topographical ROIs. Orange bars show the percentage of voxels within the ROI that had significant decoding performance. Blue bars show the subset of those voxels with a significant correlation between distance to the hyperplane and RT. These results show that decoding and RT-correlations increase in overlap from early to late areas in the ventral visual stream.

3.3 Location of the clusters with respect to the visual system

We asked in which topographical areas the clusters of significant decodable information and significant correlation were located. We determined the overlap between the decoding clusters and RT-correlation clusters with a probabilistic topographic map of visual processing areas (Wang et al., 2015). The results show an increase in overlap from early to late visual areas, with most of the overlap in areas VO and PHC in the anterior ventral visual stream (Figure 6 A&B). For the two alternative tasks on the 92-object data, we observed smaller clusters of RT-correlations (Figure 6 C&D), with relatively larger contributions from mid-level visual areas such as V3 and hV4.

3.4 Visualization of the contribution of individual exemplars to the correlations

To examine if specific exemplars (e.g., human faces) were driving the correlations, we visualized the results in representational space between distance to the classifier hyperplane and RT. At the voxel with the highest subject-average correlation, we averaged the subject-specific distances for each exemplar and displayed the pooled distance versus reaction time. We observed that for the 92-object data, human faces consistently had the largest distances and fastest RTs, and are the main source of the correlations (Figure 7 A, C, & D). When we computed the correlations for these contrasts while excluding the human face stimuli, these were not significant anymore. Importantly however, for the 118-object data, which did not have human face stimuli, frontal views of animal faces had the largest distance and shortest RTs and non-mammals consistently had shorter distances and longer RTs (Figure 7, B), showing

that the correlation between distance and RT is present in data without human faces. In sum, these results show that human faces are not required for a correlation between distance and reaction time, but when human faces are part of the stimulus set, they are the main source of the correlations.

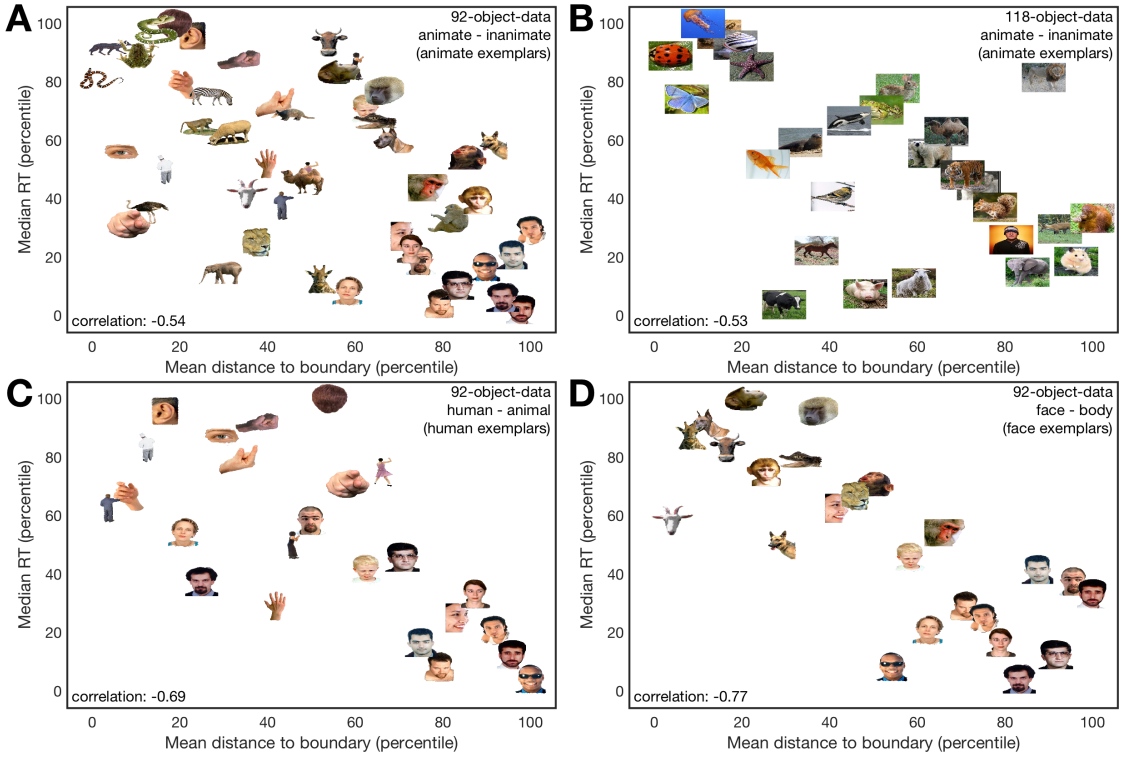

Figure 7. Scatterplots of distance to boundary versus reaction time. At the location of peak average correlation, the mean distance to the classifier boundary for each exemplar across subjects was computed. The x-axis shows the mean distance over subjects in percentiles. For each categorization task, only one side of the categorization is shown, as only one category had significant correlations with reaction time at the group level. Note that the correlations reported in the bottom left corner of the plots were between mean distance and RT, and are therefore higher than the subject-averaged correlations that were reported in the main results. Panels A,C & D show that when human faces are part of the stimulus set and the categorization task, they are consistently located in the bottom right corner, having the largest distances and shortest RTs, and are therefore be the main source of the correlations. However, in a stimulus set without human faces, such as the 118-object-data, the correlations were similar (panel B).

4 Discussion

We found that only a subset of information that is decodable could be related more directly to behaviour using the distance to bound approach, which argues for a dissociation between decodable information and information that is relevant for behaviour. In addition, we found that the distance to bound approach generalizes to new tasks and stimuli. The resulting maps of decoding and correlation with behaviour mainly corroborated the view that VTC contains behaviour-relevant object information, but also revealed correlations with behaviour in other visual areas, such as earlier in ventral stream (V3 & hV4), and in the dorsal stream (IPS).

It is important to note that with the dissociation between decodable information and information that is relevant for behaviour, we interpret the positive finding that *some* information can be related to behaviour. Crucially, this statement does not imply that the other decodable information is not used in behaviour, as this would be interpreting a null result. There are many alternative reasons for the lack of a correlation between decodable information and behaviour, which are discussed below. Our results show that a subset of decodable information can be directly related to behaviour, which challenge the common assumption in decoding studies that treats all decodable information as equally relevant for behaviour.

4.1 Dissociating between decodable information and information that is used in behaviour

Our results showed that only a subset of decodable category information correlates with independent observer categorization RTs. This shows a dissociation between information that is decodable, and information that relates more directly to behaviour. To illustrate, consider the question about what regions of the brain contribute to an animacy categorization decision (DiCarlo, Zoccolan, & Rust, 2012; Grill-Spector & Weiner, 2014). When only the result of the animacy decoding searchlight would be taken into account, this could lead to the conclusion that animacy is represented along the entire visual stream. However, when also considering the RT-correlation results, it is more evident that animacy is only suitably represented for read-out in some areas (Ritchie et al., in press; Williams et al., 2007). Most of these correlations were found in the anterior regions of the VTC, which corroborates the view that the VTC contains behaviour-relevant object information. Our findings are consistent with the ROI based approach taken in (Carlson et al., 2014), where correlations between distance and RT were observed in VTC ROIs, and to a much lesser extent in Early Visual Cortex ROIs. The searchlight approach taken here yields a localized description of the areas that are suitable for category read-out in the brain. Interestingly, areas in the prefrontal cortex were found to have decodable category information, but no correlation with RT. These areas are often considered to represent task-relevant information (Woolgar, Jackson, & Duncan, 2016). However, subjects in the fMRI experiments were performing an orthogonal task, which would explain why here prefrontal areas did not contain information that is suitable for use in decision making. It is possible that when subjects would have been performing an object categorization task in the scanner, these areas

would have represented information differently and would have shown correlations with RT.

A criticism of neuroimaging studies in general is that they only show information that is available to the experimenter, rather than information that is relevant to the brain (de-Wit et al., 2016; Ritchie et al., in press). The approach taken here illustrates this issue, as we showed that for only in a subset of the areas with decodable information, this information could be used to predict behaviour. While a correlation with behaviour still does not imply that the information is used by the brain, it provides stronger evidence for the information being used by the brain for decisions (de-Wit et al., 2016; Klein, 2016; Ritchie et al., in press; Williams et al., 2007). Here, we have shown how the distance to bound approach can be applied in a standard fMRI searchlight decoding setting to highlight decodable information that is strongly relatable to behaviour.

4.2 Limitations of the approach

A limitation of the neural distance to bound approach is that a lack of correlation between distance and RT does not imply that information is not used by the brain. When finding decodable information that does not correlate with RT, this information does not have to be irrelevant or epiphenomenal. For example, it could be that the representations are relevant for a task other than the binary categorization decision that was used here, or that other read-out models provide a better description of the process (e.g., non-linear models (Ritchie & Carlson, 2016)). However, testing all possible alternative tasks or models would be infeasible. Therefore, the distance to

bound approach only allows the positive inference on the level of suitability of decoded information in the context of a limited set of tasks, and a limited set of read-out models.

4.3 The distance to bound approach in alternative categorization tasks

Although the 118-data included more inanimate objects, there were no correlations between distance and reaction time for the inanimate stimuli. Carlson et al., (2014) argued that this effect was caused by animacy being a negatively defined category ("not animate"). Here, we tested this hypothesis using alternative categorization contrasts, namely human vs animal, and face vs body. These contrasts consisted of two positively defined categories, but as with animacy, resulted in a correlation for only one of the two ('human' and 'face', respectively). This goes against the notion of the negative definition of inanimate as main reason for a lack of correlation. However, it still is possible that observers still treated these tasks as A or NOT A, rather than A or B, while choosing 'A' based on the simplest category, as perceptual evidence for a face or human would be easier to obtain that evidence for body or animal. Thus, while not specified as a negative category, they could have been treated as such. It generally unlikely that categorization in the brain is treated as a set of binary problems. The binary categorization task is therefore possibly an unnatural way of approaching categorization, and obtaining a better description of what object categorization function is performed by the brain in natural behaviour can help direct neuroimaging research (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). This could be investigated further by including different tasks other than binary categorization, and assessing whether in those tasks the distance to bound approach provides a more complete picture of observer categorization behaviour.

4.4 Contribution of mid-level visual areas to categorization

We found that correlations with RT were not restricted to VTC, but were also prominent in V3 and hV4 (Figure 6). In the distance to bound framework, these findings suggest that representations in these areas are also suitable for read-out. Lower level features that are shared within animates and within inanimates could be a cue for read-out. For example, V4 is thought of as an intermediate stage of visual processing that aggregates lower level visual features into invariant representations (Riesenhuber & Poggio, 1999). It has been proposed that a direct pathway from V4 to areas that facilitate eye movements (e.g., FEF) accounts for fast saccadic categorization reaction times (Crouzet, Kirchner, & Thorpe, 2010; Honey, Kirchner, & VanRullen, 2008). Similar pathways to decision making areas would allow the brain to exploit visual feature cues in a fast paced animacy categorization task (Hong, Yamins, Majaj, & DiCarlo, 2016; Kirchner & Thorpe, 2006; Thorpe, Fize, & Marlot, 1996). An alternative possibility is that read out is not happening directly from V4, but its structure of the representation is shaped by the low level feature differences in animacy. This structure, then 'feeds' into more anterior areas where it remains largely preserved. This would explain similar correlations between distance to boundary and reaction time in both V4, and more anterior areas of the ventral stream. Both of these accounts are consistent with recent findings that show differential responses for object categories in mid-level visual areas based on visual feature differences, such as in animacy (Proklova et al., 2016), or object size (Long, Konkle, Cohen, & Alvarez, 2016). The extent to which shape information contributes to the read-out process could be further investigated by using the approach from this study with a stimulus set that controls for visual features (Kaiser, Azzalini, & Peelen, 2016; Proklova et al., 2016).

The similarity structures within layers of artificial deep neural networks have been shown to match similarity structures in areas in the ventral visual stream well (Cadieu et al., 2014; Güçlü & van Gerven, 2014; Khaligh-Razavi & Kriegeskorte, 2014). Eberhardt and colleagues used a similar approach as taken here and correlated distance to a hyperplane fitted on an artificial deep neural network's activations. They found that the highest correlations were obtained in intermediate layers of the deep neural networks (Eberhardt, Cader, & Serre, 2016), which is consistent with the argument that intermediate features such as shape play a major role in object categorization, and is consistent with our results of correlations between distance and reaction times in areas such as V3 and hV4.

4.5 Categorical representations in the dorsal stream

We found that information in parietal areas also correlates with RT. The classical view is that the ventral and dorsal visual streams are recruited for different tasks (Ungerleider, 1982), with the ventral stream representing object identities ('what'), and the dorsal stream spatial features ('where'). However, areas in the dorsal stream, such as IPS1 and IPS2 have been found to exhibit similar object-selective responses as areas in the ventral stream (Konen & Kastner, 2008; Sereno & Maunsell, 1998; Silver & Kastner, 2009). Information in the dorsal stream has been argued to be task-dependent, with only task-relevant information being represented in the dorsal stream (Bracci, Daniels, & op de Beeck, 2017). In our study, fMRI participants were performing an orthogonal task, making the contribution of possible task or attention to the representations minimal. However, it could be that without any visual object task, the categorical representations in the absence of a specific task resemble those found in

areas V3 and V4 of the ventral visual stream (Konen & Kastner, 2008). These representations could then be dynamically adapted to serve different tasks (Bracci et al., 2017; Freedman & Assad, 2016; Jeong & Xu, 2016).

4.6 Conclusion

In this study, we combined the distance to bound approach (Ritchie & Carlson, 2016) with a searchlight decoding analysis. We found that multiple areas in the ventral and dorsal visual streams contained decodable category information that was also suitable for read out in behaviour, as distance to classifier boundaries obtained from these areas correlated with observer categorization reaction times. These correlations support the large role VTC plays in object categorization. However, they also suggest that mid-level ventral and dorsal areas contribute to categorization decisions. Our results speak to the current debate in Neuroimaging research about whether information that we can decode is the same information that is used by the brain in behaviour (de-Wit et al., 2016). With our approach, we highlighted that decodable information is not always equally relevant for the brain in behaviour.

Acknowledgements

This research was supported by an Australian Research Council Future Fellowship (FT120100816) and an Australian Research Council Discovery project (DP160101300) awarded to T.A.C., and a German Research Foundation grant (CI-241/1-1) awarded to R.M.C.

References

Ashby, F. G. (2000). A Stochastic Version of General Recognition Theory. *Journal of Mathematical Psychology*, *44*(2), 310–329. https://doi.org/10.1006/jmps.1998.1249

Ashby, F. G., & Maddox, W. T. (1994). A Response Time Theory of Separability and Integrality in Speeded Classification. *Journal of Mathematical Psychology*, *38*(4), 423–466. https://doi.org/10.1006/jmps.1994.1032

Bracci, S., Daniels, N., & op de Beeck, H. (2017). Task Context Overrules Object- and Category-Related Representational Content in the Human Parietal Cortex. *Cerebral Cortex*, 1–12. https://doi.org/10.1093/cercor/bhw419

Bracci, S., & op de Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. *Journal of Neuroscience*, *36*(2), 432–444. https://doi.org/10.1523/JNEUROSCI.2314-15.2016

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Caramazza, A., & Shelton, J. R. (1998). Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience*, *10*(1), 1–34. https://doi.org/10.1162/089892998563752

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction Time for Object Categorization Is Predicted by Representational Distance.

Journal of Cognitive Neuroscience, *26*(1), 132–142. https://doi.org/10.1162/jocn_a_00476

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of Activity in the Categorical Representations of Objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. https://doi.org/10.1162/jocn.2003.15.5.704

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1. https://doi.org/10.1167/13.10.1

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, *20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/10.1016/S1053-8119(03)00049-1

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4), 16. https://doi.org/10.1167/10.4.16

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*(5), 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Downing, P. E., Chan, A. W.-Y., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain Specificity in Visual Cortex. *Cerebral Cortex*, *16*(10), 1453–1461. https://doi.org/10.1093/cercor/bhj086

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414

Downing, P. E., & Peelen, M. V. (2016). Body selectivity in occipitotemporal cortex: Causal evidence. *Neuropsychologia*, *83*, 138–148. https://doi.org/10.1016/j.neuropsychologia.2015.05.033

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How Deep is the Feature Analysis underlying Rapid Visual Categorization? In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 1100–1108). Curran Associates, Inc.

Freedman, D. J., & Assad, J. A. (2016). Neuronal Mechanisms of Visual Categorization: An Abstract View on Decision Making. *Annual Review of Neuroscience*, *39*(1), 129–147. https://doi.org/10.1146/annurev-neuro-071714-033919

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *New York*.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Güçlü, U., & van Gerven, M. A. J. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. *PLoS Comput Biol*, *10*(8), e1003724. https://doi.org/10.1371/journal.pcbi.1003724

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, *17*(4), 323–328. https://doi.org/10.1016/j.cub.2006.11.072

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision*, *8*(12), 9–9. https://doi.org/10.1167/8.12.9

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247

Jeong, S. K., & Xu, Y. (2016). Behaviorally Relevant Abstract Object Identity Representation in the Human Parietal Cortex. *Journal of Neuroscience*, *36*(5), 1607–1619. https://doi.org/10.1523/JNEUROSCI.1016-15.2016

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology*, *115*(4), 2246–2250. https://doi.org/10.1152/jn.01074.2015

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Comput Biol*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309. https://doi.org/10.1152/jn.00024.2007

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776. https://doi.org/10.1016/j.visres.2005.10.002

Klein, C. (2016). Brain regions as difference-makers. *Philosophical Psychology*, *0*(0), 1–14. https://doi.org/10.1080/09515089.2016.1253053

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, *11*(2), 224–231. https://doi.org/10.1038/nn2036

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00128

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, 27. https://doi.org/10.3389/fninf.2016.00027

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially Distributed Representations of Objects and Faces in Ventral Temporal Cortex. *Journal of Cognitive Neuroscience*, *17*(4), 580–590. https://doi.org/10.1162/0898929053467550

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/jocn_a_00924

Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2010). Quantifying the Adequacy of Neural Representations for a Cross-Language Phonetic Discrimination Task: Prediction of Individual Differences. *Cerebral Cortex*, *20*(1), 1–12. https://doi.org/10.1093/cercor/bhp076

Redcay, E., & Carlson, T. A. (2015). Rapid neural discrimination of communicative gestures. *Social Cognitive and Affective Neuroscience*, *10*(4), 545–551. https://doi.org/10.1093/scan/nsu089

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and “Inner” Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience*, 190. https://doi.org/10.3389/fnins.2016.00190

Ritchie, J. B., Kaplan, D. M., & Klein, C. (in press). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLoS Comput Biol*, *11*(6), e1004316. https://doi.org/10.1371/journal.pcbi.1004316

Sereno, A. B., & Maunsell, J. H. R. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature*, *395*(6701), 500–503. https://doi.org/10.1038/26752

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678. https://doi.org/10.1162/jocn_a_00733

Silver, M. A., & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences*, *13*(11), 488–495. https://doi.org/10.1016/j.tics.2009.08.005

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509. https://doi.org/10.1146/annurev-psych-120710-100412

Ungerleider, L. G. (1982). Two cortical visual systems. *Analysis of Visual Behavior*, 549–586.

van Bergen, R. S., Ji Ma, W., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, *29*(34), 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage*, *132*, 59–70. https://doi.org/10.1016/j.neuroimage.2016.02.019

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*(6), 685–686. https://doi.org/10.1038/nn1900

Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. *Journal of Cognitive Neuroscience*, *28*(10), 1433–1454.

Chapter 5

Discussion

Brain decoding methods have become standard practice in analysing fMRI data and, more recently, MEG data. A current debate is whether information that is extracted with brain decoding methods is the same information that is relevant for the brain, for example, whether it is used in behaviour. The aim of this thesis was to explore methods for decoding brain representations and linking them to behaviour. This consisted of first empirically assessing options for the novel application of decoding methods to neuroimaging data with high-temporal resolution, such as MEG. Secondly, the recently proposed neural distance to bound approach was explored as a method for linking brain decoding methods to behaviour.

In the previous three chapters, I presented the empirical findings of this thesis. In chapter two, we used example MEG data to illustrate a typical MEG decoding analysis and compared the effect of different options along the pipeline. The results showed that these choices can affect local statistical significance of the results, and would thus potentially impact the conclusions. Chapter two focused on developing decoding methods, which were built upon in the rest of the thesis. In chapter three, I used MEG decoding in combination with the neural distance to bound approach (Ritchie and Carlson, 2016). I tested whether this approach follows the prediction of a behavioural manipulation. The results showed that the distance to the classifiers hyperplane successfully predicted

reaction time, accuracy, and the parameters of a model of the observer decision process. The results further showed that the distance to bound approach could only account for the effects on animate objects. In chapter four, I further examined the relationship between decodable information and behavior by creating spatially unbiased maps of where the distance to the classifier hyperplane could be used to predict behavior. Using a searchlight approach on fMRI data, classifiers were first trained to decode object category information in local voxel clusters. Secondly, the distance to bound approach was used to test whether this information can also be used to predict categorization reaction times. The results show that decodable information exists along the entire ventral and dorsal visual streams, but that behavior can only be predicted from a subset of those locations.

The results presented in this thesis support the argument that significant decoding is not enough to show the presence of information that is relevant for the brain (de-Wit et al., 2016; Ritchie et al., in press). The findings in chapter two show that subtle choices in the analysis pipeline can affect the strength of decoding, and can for example move the onset of significant decoding. While this study mainly focused on developing the MEG decoding methods, the findings do emphasise that significant decoding results should be interpreted with appropriate caution. Chapter three showed that the onset and peak time points of the decoding time series differ from the onsets and peaks of the time-varying correlations with behaviour. This difference was even greater with the correlation to drift rate. This suggests that the time of peak decoding is not necessarily the time that best predicts behaviour (cf. Ritchie et al., 2015). Chapter four mapped the dissociation between decodable information and information that can be read out in behaviour, and showed that not all decodable information is equally suitable for read out (assuming a

linear readout model). Taken together, the results in this thesis show that observing decodable information does not imply that this information is being used by the brain.

The experiments in chapters three and four used the distance to bound approach as a method to test whether information can be used by the brain in behavioural read out. To date, this method has been applied in the context of the same stimulus set and one task (Carlson, 2014; Ritchie et al., 2015). Therefore, limited conclusions can be drawn about the generalizability and robustness of the distance to bound approach. This was addressed in in chapters three and four by using novel stimulus sets and alternative tasks, in both MEG (chapter three) and fMRI (chapter four). The results presented in those chapters showed that the approach can account for the effect of degrading (chapter three) and that it generalizes to different tasks (chapter four) and stimuli (chapter three and four). In addition, the results showed how distance to the classifier boundary can be directly related to evidence accumulation models of observer decision making behaviour (chapter three). These results show that the distance to bound approach is a useful method to dissociate decodable information from information that is relevant for behaviour, and how this dissociation can be mapped out in space and time.

In the current chapter, I will discuss the implications of these findings and how they contribute to the current literature. As this thesis made use of relatively novel MEG decoding techniques, I will first discuss how the findings in this thesis contribute to the development of MEG decoding (section 1). I then move on to the second main focus of the thesis, which is how to go beyond decoding with the distance to bound approach, and outline how the approach can help with interpreting decoding results (section 2). I will then review how my results are in support of the general distance to bound approach

(section 3), but also what limitations and questions on the specific workings on the distance to bound approach have been raised in my research (section 4). Next, as visual object categorization was used as the domain in which to test the distance to bound methods, I relate my results to the visual object categorization literature (section 5). Finally, I reflect on this thesis in light of two recent articles that present critical views on the general approaches taken in Cognitive Neuroscience and present my broad views on the future directions of the field (section 6), before bringing everything together with a general conclusion (section 7).

1 MEG decoding

MEG decoding is a promising field, and the relatively small number of MEG-decoding studies conducted to date have provided many valuable insights (Contini et al., in press; Varoquaux and Thirion, 2014). MEG has many advantages over EEG as method for studying the temporal dynamics of information with millisecond accuracy, such as fewer artefacts, less signal smearing and distortion, and better signal source approximations (Baillet, 2017). A wide array of toolboxes implementing MEG preprocessing and decoding are now available (Baillet et al., 2011; Delorme and Makeig, 2004; Hanke et al., 2009a; Oostenveld et al., 2010; Oosterhof et al., 2016; Tadel et al., 2011; Varoquaux and Thirion, 2014). However, these methods are still under active development. Problems with current analysis methods continue to arise (e.g., Allefeld et al., 2016; Eklund et al., 2016; Haufe et al., 2014; Ritchie et al., in press; Thirion et al., 2015), and their solutions need to find their way into the toolboxes and general literature. For example, a commonly applied (mostly exploratory) analysis was to project the classifier's weights back onto the channels after a decoding analysis to investigate the source of the

decodable signal (see e.g., Carlson et al., 2013; Chan et al., 2010). However, popular classifiers such as LDA and SVM take into account the covariance between features. Because of this, their weight vectors can therefore not be directly interpreted (Haufe et al., 2014). Haufe et al., (2014) present a solution in the form of a transformation from weights to activation patterns. The transformation has been incorporated in fMRI decoding toolboxes (Hebart et al., 2015), and the issue is described in recent decoding reviews (e.g., Haynes, 2015), including chapter two of this thesis for its application to time-series studies. This helped with creating a general awareness about such caveats and ensured inclusion of the weight transformation method in recent decoding studies using fMRI (e.g., Ritter et al., 2014; Wardle et al., 2017) and MEG (e.g., Wardle et al., 2016).

Taken together, the number of studies that use MEG decoding methods is rapidly increasing (Contini et al., in press; Varoquaux and Thirion, 2014). While toolboxes for MEG decoding now exist (Gramfort et al., 2013, 2014, Hanke et al., 2009a, 2009b; Oosterhof et al., 2016; Tadel et al., 2011), these still require some background in machine learning and programming experience (Varoquaux and Thirion, 2014). Therefore, it is important to continue efforts on describing, testing, and developing decoding methods, as was done in chapter two of this thesis. While the aim of chapter two was not to present a definitive guideline to performing decoding analyses, it provides a useful guide for researchers who are setting up their own decoding pipelines and additionally raises awareness about potential issues and caveats specific to decoding analyses. Moreover, continuing to test decoding methods in different scenarios is important to detect interactions and possible caveats in for example preprocessing and feature selection, and choice of classifier (Chapter two). The research presented in

chapter two of this thesis involved substantial experimentation with MEG decoding. In the next sections, I will reflect on how this has led to general thoughts on the directions of the current MEG decoding literature. I will argue that the pipelines used in univariate analysis do not necessarily apply to MEG decoding, and that the aims of MEG decoding studies should in general not focus heavily on spatial inferences.

1.1 Revisiting traditional analysis pipelines

Chapter two and three of this thesis used MEG decoding to reveal the temporal dynamics of information in a signal with high temporal resolution. A methodological point made by these studies is the fact that good decoding performance can be obtained with a relatively limited amount of preprocessing and using the simplest classifiers. Compared to the traditional MEG component analysis (cf. Luck, 2005), several commonly performed steps were skipped. Steps such as eye-blink removal, baseline correction, and filtering are part of most preprocessing pipelines. The findings in chapters two and three argue against the need for many of these procedures for conducting decoding experiments. An argument for not performing standard preprocessing steps is that classifiers can handle classic sources of noise in the data. For example, consider a bad channel that only contains only noise. In an ERP analysis, including this channel will result in a noisier average signal, and it should therefore be excluded from the analysis. In a decoding analysis, the classifiers will in their training stage find that the means and variances of the conditions are equal in this channel, and thus assign it a zero-weight. Therefore, there is no reason to exclude the channel, as resulting decoding performance of the classifier is likely to be unaffected.

Including a noisy channel can also aid decoding; consider a second channel that measured the same noise, but also has a small underlying signal that differentiates the conditions (Haufe et al., 2014). This underlying signal can be recovered by subtracting the first noisy channel from the second, thus both channels are 'used' by the classifier. Similarly, other sources of noise in ERP components such as eye-blinks, or noise at higher frequencies do not affect decoding performances in the same way, and it is therefore not needed to perform all the standard steps. Instead, MEG decoding analyses should have their own separate pipelines, aimed and tailored to the classifiers. For example, even if some trials contain eye blinks or other noise, a balanced number of trials per condition is arguably more important for classification, to avoid biased classifiers. Finally, reducing the number of preprocessing steps results in fewer combinations of free parameters in the design that can lead to false positives (Carp, 2012; Poldrack et al., 2017; Simmons et al., 2011; Strother, 2006; Varoquaux et al., 2016). In sum, despite the established protocols for univariate analyses, MEG decoding studies should not by default adhere to these pipelines, and instead develop in a separate direction.

1.2 Spatial inferences in decoding studies

A common aspect of ERP analyses is to compute the ERPs at each individual channel, to find the location of the differences in signal between conditions. It is often of interest to recover the source channel(s) of information in a decoding setting, by examining the weights the classifier assigned to each channel. However, as described in chapter two, the interpretation of this is not trivial for two reasons. First, the weights are only as good as the classifiers performance. If the classifier obtains a low decoding performance, its

weights are not likely to reflect the source of the information. Secondly, the weights reflect covariance in the data. As such, a high weight does not imply a large difference in signal, and the weights need to be transformed before interpreting (Haufe et al., 2014). However, after transforming, the resulting pattern differences will be similar to a mass univariate GLM (Haufe et al., 2014), and thus do not yield more insights than such an analysis.

Because of these difficulties in interpreting classifier weight vectors, the usefulness of creating topographical maps in a decoding analysis is debatable. An approach for spatial localization that is more viable, is to use a sensor searchlight (e.g., Kaiser et al., 2016a). This involves repeating the analysis for one sensor and its direct neighbours. If the local group of sensors contain information, it will lead to significant decoding accuracies. These are then stored at the centre sensor, to allow comparing decoding accuracies across topographical maps. Alternatively, the MEG data can first be warped into a virtual source space (Van Veen et al., 1997), and the classification can be performed in regions in source space. However, the results would highly depend on the quality of the reconstruction at the single trial level. The algorithms for source estimation are still under active development, and it is therefore likely that reconstruction accuracies will improve in the future (Baillet et al., 2011; Baillet, 2017). In sum, MEG decoding approaches are currently limited in the spatial inference that can be made, and should therefore mainly focus on temporal dynamics.

With the strength of MEG decoding lying in the temporal dynamics, it can be combined with fMRI to yield a full picture of the spatio-temporal aspects of emerging information. Cichy et al., (2016, 2014) demonstrated this using representational similarity analysis to

show where in time and space the MEG and fMRI signals contained similar structure during object recognition (Cichy et al., 2014, 2016). Future research efforts could improve on this approach by modelling explicit categorical information in the overlapping structures. For example, as was done in Carlson et al., (2014), the shared RDM can be used to reconstruct virtual representational dimensions on which an animacy classifier can be trained that would yield a measure of whether animacy information is available in the shared RDM. This classifier would allow applying the distance to bound method to show *where* and *when* behavioural readout is possible in the shared fMRI-MEG representations. This combination of methods would then provide a complete description of the spatio-temporal dynamics of category information read-out by the brain.

2 On interpreting decoding results

The interpretation of decoding studies is not straightforward. As discussed in Chapter one of this thesis, a common assumption that is often made in decoding studies is that if information is available to the researcher, then it is also available to the brain to use (de-Wit et al., 2016; Ritchie et al., in press). Even though most would agree that this assumption is false, it is often not discussed or mentioned as a caveat in decoding studies. The next section discusses how this important theoretical issue argues for a shift in perspective and highlights the need for the development of methods to address the link between brain and behaviour.

2.1 Decoding does not measure information

By asking what constitutes ‘information’, de-Wit et al., (2016) argue that neuroimaging research efforts should shift their focus away from testing the availability of information to the experimenter. Instead, information has to be shown to be available to the brain (de-Wit et al., 2016). The two types of information are not equal. For example, classifiers in a decoding analysis are able to decode category information from early visual areas (Williams et al., 2007), even though these areas are not connected to decision making areas and are in general not believed to explicitly represent categories (DiCarlo and Cox, 2007; Grill-Spector and Weiner, 2014; Kravitz et al., 2013). In this case, it is unlikely that the category information in early visual areas is used by the brain in making a categorical decision (de-Wit et al., 2016; Williams et al., 2007). Therefore, de-Wit et al. argue, it is important to be clear about what to define as information in the brain. In neuroimaging studies, information should only be called ‘information’ if it can be shown to be available to the brain itself, but current MVPA techniques can only show information that is available to the experimenter (de-Wit et al., 2016). The term information is currently used more broadly in the literature, and also at various locations in this thesis, to refer to information that is available to the experimenter. In contrast, de-Wit et al. argue that the term 'information' must be used with caution in communicating neuroimaging results to avoid provoking claims such as finding ‘the neural representation of objects’. However, theoretical shifts like that will have to come slowly and gradually. For example, the field is currently shifting away from using similarly misleading terms such as “brain reading”, which were commonly used in the earlier fMRI decoding literature (Cox and Savoy, 2003; Haynes et al., 2007; Norman et al., 2006).

In chapter four, using the distance to bound approach showed the distinction between areas with decodable information (with “experimenter as receiver” (de-Wit et al., 2016)), and areas where the decodable information also correlated to categorization behaviour on an individual exemplar basis. Here, the latter reflects information that is available to the brain for read-out in behaviour (with “cortex as receiver” (de-Wit et al., 2016)). Being able to model the read-out of information can be taken as evidence for the information being available to the brain as receiver (de-Wit et al., 2016; Ritchie et al., in press). The information in chapter four is thus defined as object animacy, which we assumed the brain reads out, and uses to make the animacy categorization decision. However, it still does not answer if and how the information that did not correlate to behaviour is used by the brain. Moreover, relating distance to boundary with reaction times constitutes a large jump, and critically misses the processes in the brain that transform the information into a decision. Therefore, it can still not provide a complete picture of the process. Currently however, no definitive methods exist that show how information is used by the brain (as also argued in de-Wit et al., 2016) which highlights the need to further develop approaches such as the distance to bound approach.

3 On the distance to bound approach

This thesis extensively experimented with using the distance to bound approach as a method to relate decoded information to behavioural readout. A limitation of the research using this method to date, is it was all conducted on only one stimulus set and categorization task (animacy). In the experiments presented in chapters three and four, I found that the distance to bound approach generalizes to other sets of stimuli and categorization tasks (e.g., faces vs bodies and humans vs animals). The new stimuli

included backgrounds, instead of isolated objects, which in general worsens decoding accuracies (Coutanche et al., 2016). The backgrounds allowed to better control for shape, which could interact with the classifier (and therefore with the correlations). In addition, in chapter three, short stimulus durations were used (66ms) compared to the relatively long 500ms used in Ritchie et al., (2015). While less exposure to the stimulus would affect the dynamics of the MEG signal, the resulting correlations between distance and reaction time were strikingly similar suggesting that stimulus duration does not affect the dynamic read out. These results show that, at least for animacy categorization, the distance to bound approach is robust to different types of stimuli and experimental set-ups. Most notably, distance correlated with reaction time in the absence of face stimuli in chapters three and four. The previously reported correlations (Carlson et al., 2014; Ritchie et al., 2015) could have largely been driven by the face stimuli, as faces evoke fast reaction times (Crouzet et al., 2010; Crouzet and Thorpe, 2011) and distinct patterns of activation (Haxby et al., 2001; Kanwisher et al., 1997; Kriegeskorte et al., 2008). When using the same stimulus set in chapter four, this was also observed. Yet, when using stimulus sets that did not include human faces, the correlations were similar to those previously reported, both in MEG and fMRI. Taken together, this shows that face stimuli are not necessary for a correlation between distance and reaction time.

An important result was the observation that distance to the classifier boundary correlates with drift rates estimated using LBA. This result shows that the distance to bound approach can be related to current evidence accumulation models of decision behaviour, which is an important step towards developing a linking approach between brain and behaviour. As LBA is a flexible tool to model many different tasks, it allows us to go beyond using reaction times as a dependent variable and proxy for decision

difficulty, and link neural spaces to more complete models of behaviour (Forstmann and Wagenmakers, 2015; Purcell et al., 2010; Purcell and Palmeri, 2016). However, the correlation between distance and drift rate was not significantly higher than distance and RT. This suggests that, at least for the binary animacy categorization task, enough of the difficulty is equally well captured by RT, or that the additional variance captured by drift rate is obscured by the noise in the neuroimaging data. It remains to be shown that when using other tasks that for example affect accuracy more than RT, drift rate correlations would outperform those for RT. Regardless, the correlation with drift rate supports the biological plausibility of classifier hyperplanes as linear decision boundaries in neural spaces (DiCarlo and Cox, 2007; Ritchie and Carlson, 2016).

An outstanding question is on how to compute the decision boundary. To date, it has been fitted to activation patterns, to simulate a decision boundary that the brain could use. However, to emulate the whole read out process, accumulation processes must be modelled dynamically, with evidence gradually increasing. Previous research has for example used the rising strength of ERP components as evidence in an accumulation model (Bennett et al., 2015; Philiastides and Sajda, 2006). A threshold should be included to supress baseline activity from contributing to evidence accumulation. The temporal nature of MEG decoding is optimally suited for this purpose. In addition, MEG can be used to establish at what time evidence accumulation would start. Ritchie et al., (2015) argued that the read out process would follow the availability of information and would therefore match the decoding trace. Contrastingly, the results from chapter three showed that the peak correlations occurred before the time point of peak decoding, which suggests that much of the read out process occurs before the time point of optimal decoding. To investigate this further, future research could more closely test different

read out strategies that for example emphasise speed (Rae et al., 2014), and see how these relate to decoding performances.

4 Limitations of the distance to bound approach

The findings in this thesis highlight limitations of the distance to bound approach, that need to be addressed in order to make the approach a viable option for linking brain decoding methods to decision behaviour. This section discusses these limitations in more detail. A critical issue with the distance to bound approach is that only positive correlations can be interpreted. A lack of a correlation between distance and RT is effectively a null-result, and thus cannot be taken to mean that information is not used by the brain. In the case of no correlation, it is unclear whether this is due to a lack of power in the decoding or RT measurements. It could also be possible that the wrong behavioural task is used. Another option is that the choice of stimuli led to category decoding performance (in irrelevant brain areas) without RT-correlations because of confounding low level visual features. By controlling such confounds with carefully selected stimuli (e.g., Bracci and op de Beeck, 2016; Proklova et al., 2016) or using cross-decoding approaches (e.g., Kaiser et al., 2016a, 2016b), these explanations can be ruled out.

The studies to date, including this thesis, have found correlations for only one side of the categorization. The reason for this is still unclear, and a challenge for future research is to find a solution for read out of both sides of the categorization that would provide a more compelling model of linking brain decoding to categorization decision behaviour. Finally, while correlations between distance and RT provide evidence for information

being suitably formatted to be used by the brain, the approach is still correlational in its nature. In sum, the main limitation of the distance to bound approach is that it is only a step towards modelling the brain-behaviour relationship. To properly address the question whether decoded information is used by the brain, I argue that a more complete solution is needed.

4.1 Evidence from correlations

The distance to bound approach relies on correlations between brain and behaviour to obtain evidence for read-out. However, as is the case with all correlational research, a positive correlation does of course not imply a causal relationship. An unknown co-varying factor could be the true source of the relationship. However, as also argued in de-Wit et al., (2016), while we still need to focus on finding ways of exploring causation, including behavioural correlations in decoding research is a good practice. In contrast, Klein (2016) argues that strong evidence for causation can be inferred from finding systematic relationships. A currently highly relevant example is the systematic relationship between the amount of human-produced greenhouse gasses emitted into the atmosphere and the average global temperature, which is accepted as a causal relationship in the current scientific consensus (IPCC, 2013). Klein (2016) similarly considers the scenario of a highly systematic relationship between a neural variable and behaviour, where from a change in the neural variable one can predict the change in behaviour. If this relationship is specific and systematic enough, it would then provide compelling evidence for a causal relationship (Klein, 2016).

Previous research has correlated decoding accuracies to behavioural accuracies, and found those correlations only in a subset of areas that had decodable information (e.g., Williams et al., 2007), which provides more evidence for one area being involved in the read-out. Other approaches have correlated decoded similarities with behavioural similarities, such as differences in reaction time (e.g., Mohan and Arun, 2012), or perceptual similarity (e.g., Proklova et al., 2016; Wardle et al., 2016). These approaches all address the question of relating decoded information to behaviour, and are therefore similar to the distance to bound approach. However, the distance to bound approach explicitly models the read-out of individual exemplars rather than correlating mean performances. Therefore, a correlation between individual exemplar distances and reaction times provides even more compelling evidence that a representation is suitable for read-out in behaviour (Ritchie et al., in press). Even though it does not imply a causal relationship, it is more compelling evidence to be able to show a systematic relationship between brain decoding at the exemplar level and behaviour. Even stronger evidence for a systematic relationship comes from the findings in chapter three, where manipulating one variable (behaviour) resulted in a matching change in the other (distance). Therefore, the distance to bound approach is a good first step towards revealing a highly systematic relationship which in the framework of Klein (2016) can be attributed to causation. Following that, the distance to bound approach therefore has the potential to show information in the brain, in the framework of de-Wit et al., (2016). However, there is still a long way to go, and some serious limitations and outstanding questions have to be considered. In the following sections, these are reviewed in detail.

4.2 Explaining asymmetric correlations in categorization

Chapters three and four did not find correlations for inanimate stimuli, consistent with previous research. The lack of inanimate correlations was previously argued to be caused by inanimate being defined as a negative category (not animate). However, this does not explain why in chapter three, no effects of degrading were observed for the inanimate stimuli. Moreover, in chapter four, alternative categorization tasks such as faces vs bodies and humans vs animals resulted in correlations for only one of the two categories (faces and humans, respectively). These findings suggest that the negative definition of inanimate is not the reason for a lack of two-sided correlations. An alternative explanation is that the two-sided categorization task is not well suited in this setting. It is possible that, instead of a true binary categorization, observers simply treat it as an X-detection task with X being the easiest category to obtain evidence for. Many animals share features, such as shape, and therefore evidence for 'animal' is easier to obtain than for all possible inanimate objects, which do not share specific features (Caramazza and Shelton, 1998; Rosch et al., 1976). The same strategy could apply in a face vs body task, where faces have only a very specific shape and configuration and thus evidence for faces is easier to obtain than that for bodies. For humans and animals this would predict that humans are the easier category with similar shapes compared to a larger set of possible animal options to consider. The correlations for faces and humans and lack of correlations for bodies and animals in chapter four are consisted with this. Taken together, in the research to date, correlations have been driven by one side of the categorization. The results from using different tasks show that the distance to bound approach fails when modelling both sides of the categorization at the same time, even when both are defined positive.

4.3 Modelling alternative tasks

The lack of correlations for one side of the categorization suggest that binary categorization tasks might not be suitable in this context. It is possible that the representation is used in different tasks, such as detection (go/no-go, e.g., Macé et al., 2009; Mack et al., 2008), object naming (e.g., Bruner and Potter, 1964), categorical search (Maxfield et al., 2014; Zelinsky et al., 2013), or 2AFC paradigms (e.g., Wu et al., 2015). An issue with these tasks is that it would often be unclear where to put the neural decision boundary, as MVPA classification is inherently binary. To tackle these tasks, other classification approaches could be considered, such as one class classification (Khan and Madden, 2014; Minter, 1975) for a go/no-go task. To date, the distance to bound approach has only been considered in binary categorization tasks, and more studies are needed to investigate how it works in other task settings. Moreover, the results to date have been shown to be independent of the task performed in the scanner. This is surprising, considering that representational spaces and decoding time-courses have been found to be significantly affected by the task (Harel et al., 2014) and attentional set (Çukur et al., 2013; Kaiser et al., 2016b; Kay et al., 2015; Nastase et al., 2017). Future research can explore whether such manipulations predict the corresponding changes in observer behaviour, using a similar approach as was taken in Chapter three of this thesis.

A fruitful avenue to pursue for this goal is to use evidence accumulation models fitted to various tasks and predict their drift rates. As shown in chapter three of this thesis, drift rate correlates with distance equally well as reaction time. Evidence accumulation models have been adapted to other tasks, for example in a go/no-go task, LBA has been

shown to fit reaction times well (Brown and Heathcote, 2008; Rae et al., 2014). These LBA versions use two accumulators, one for correct and one for incorrect 'go' responses (Rae et al., 2014). The drift rates of these accumulators could then be correlated to binary classifiers that predict go/no-go responses.

4.4 Distance to bound in other domains

Aside from the lack of tests using different tasks, the distance to bound approach has also only been tested on categorizing visual object stimuli. In order to fully assess the ability of the approach to link brain and behaviour, it would have to be tested on different types of stimuli. In chapter four it was shown that the higher object processing areas have the strongest correlations between distance and reaction times for an object categorization task, which matches the idea of those areas containing the abstract categorically structured object representations (Carlson et al., 2014). This reasoning predicts that for a different type of stimulus, for example when categorizing the orientation of stimuli, the highest correlations would be found in early visual areas. Another prediction would be categorizing the direction of movement, which is thought to be best represented by areas V5/MT. To fully encompass the suitability of the distance to bound approach, it has to be generalizable to these other domains.

5 Visual object categorization

This thesis mainly addresses the interpretation of decoding results and moving towards linking decoding methods to behaviour. Object categorization was used as a domain to test and develop these methods on. Hence, the results also provided insights into visual object categorization.

5.1 The animacy distinction

The experiments in chapter three and four used the robust animacy distinction to test the distance to bound approach, and to map out the dissociation between decodable information and information that relates to behaviour. In these experiments, the correlations were always driven by the animate stimuli, which is consistent with results from Carlson et al., (2014) and Ritchie et al., (2015). Surprisingly, when experimentally manipulating the difficulty of the stimuli in chapter three, animate stimuli moved closer to the boundary, but inanimate stimuli did not. These findings support the notion that animate and inanimate are not equivalent categories (Caramazza and Mahon, 2003; Caramazza and Shelton, 1998). It is possible that when using animacy as top level categorization, the inanimate category should be defined at the basic level instead. Thus for animates, the categorization should be made against e.g., tools, vehicles, or fruit. However, these categories are not as strongly decodable from neuroimaging data compared to animacy, and it is therefore not clear whether a good estimate of the decision boundary can be made for such comparisons.

A somewhat surprising observation using the distance to bound approach is that it does not seem to make a difference whether the reaction times were collected using the same subjects as the neuroimaging data. In Carlson et al., (2014), a reconstructed neural space was used to fit the classifier and obtain the distances, while reaction times were collected on Amazon's Mechanical Turk. In Ritchie et al., (2015), the reaction times were collected simultaneously with the neural data. The approaches yielded comparable results. More strikingly, Ritchie et al., (2015) found similar correlations regardless of whether participants were actively categorizing, or performing a distractor task. The results in chapter three of this thesis were also obtained using behavioural and neural data from the same participants, while chapter four combined fMRI data with behavioural data collected on Mechanical Turk. A correlation between distance to boundary and reaction time for animacy categorization was found in all experiments. This shows that the animacy distinction in the human brain is very similar between humans, and suggests that individual visual experience plays a minor role in shaping the boundary (cf. Caramazza and Mahon, 2003; Haxby et al., 2001; Mahon et al., 2007; Rogers et al., 2005; Tarr and Gauthier, 2000). Moreover, studies have shown distinctive categorical responses to spoken words in the ventral visual pathway in congenitally blind participants (Bi et al., 2016; Mahon et al., 2009; Peelen et al., 2013, 2014; Ricciardi et al., 2014; Wang et al., 2017). If these results have a similar fine-grained underlying organization as evoked by visual stimuli, future studies can test whether the distance to bound results from chapter four generalize to other modalities, such as spoken words.

5.2 The contribution of visual features to categorization

In chapter four, the areas that had a significant correlation between distance and reaction time were mapped in the brain. These maps showed correlations in lower level visual areas, such as V3 and hV4, in addition to the higher (anterior) areas in VTC. This result suggests that these areas can contribute to the animacy read out. A currently debated issue is whether mid-level features contribute to the animacy organization in the ventral stream (Coggan et al., 2016; Gaspar and Rousselet, 2009; Long et al., 2016; Proklova et al., 2016; Ritchie et al., in press). The correlations in earlier visual areas speak to this, as they suggest that when making animacy decisions, the brain could use information from these areas and could thus rely on exploiting mid-level features. A question that remains unanswered is how to show whether the mid-level visual areas are used in read out, or whether their apparent organization by animacy is a by-product of animals sharing certain features. The results from using the distance to bound approach in chapter four cannot be used to distinguish between these two options, and to do so, causal approaches might be required.

6 Future challenges in Cognitive Neuroscience

As discussed in section 2 of this chapter, the interpretation of decoding results is a fundamental issue in Cognitive Neuroscience, and new methods are needed to help the interpretability of neuroimaging studies. Another ongoing debate is to what level (Cognitive) Neuroscience can lead to an understanding of the brain (Figure 1). Marr argued that to understand a cognitive function, it needs to be described at three levels (Marr, 1982). The top level is the computational level, which defines the global input-output function of the process under study. The second level is the algorithmic level, which describes how the input-output function is performed. The third and final level is the implementation level, which characterizes the underlying hardware that performs the algorithms. As neuroscience focuses mainly on the implementation level, it is currently debated whether the algorithmic level can be inferred from the implementation. In this section, I will discuss the recent article by Krakauer et al., (2017), who argue that neuroscience needs to incorporate behavioural studies more to come to an understanding of the algorithmic level of cognitive functions.

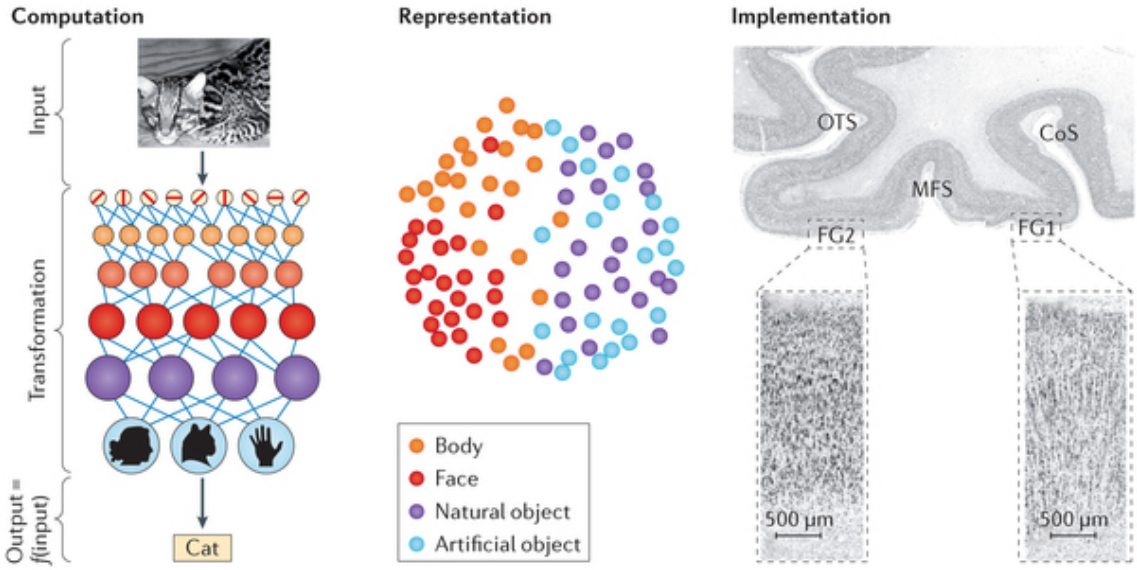

Figure 1. Visual object categorization at Marr's three levels of analysis. At the computational level, the input and output relationship is described. The algorithmic (here named representation) level describes how these computations are performed, that is, the algorithms and types of representations that are used to transform the input into the output. Finally, the implementation level describes how the algorithms and representations are implemented by the physical properties of the brain. A complete understanding of visual object categorization requires a description at all three levels (Marr, 1982). Figure from Grill-Spector and Weiner, (2014).

6.1 Combining decoding with behaviour

Krakauer and colleagues argue that neuroscience results are not useful in isolation and that a more careful study of the behaviour that is to be explained is needed (Krakauer et al., 2017). They argue that often, non-ecologically valid behaviours are studied, which are not relevant for understanding natural behaviours. More importantly, they argue that studying the neural system can only describe the brain at Marr's implementation level (Marr, 1982). According to Krakauer et al., a description at Marr's algorithmic level is needed first, which a detailed analysis of behaviour can provide (Krakauer et al., 2017). Currently, the neuronal (implementation) level is the focus of a large body of work, and careful specification of the cognitive function under study is rarely included in the studies (Krakauer et al., 2017).

Conducting behavioural experiments to guide neuroscience research is important, and there is currently limited inter-play between neuroscience and behavioural research. Moreover, investigating possible sources of a behavioural effect with neuroimaging can be good ways to test specific predictions. For example, in chapter three of this thesis, we tested the hypothesis that a shorter neural distance to bound can explain the decrease in behavioural performance for noisy stimuli. Here, the experiment was designed around explaining a behavioural effect. A more complete investigation could first specify the exact cognitive function that is of interest (Krakauer et al., 2017; Marr, 1982), describing what its inputs and outputs are (the computational level). Then, by constructing theories about the underlying algorithms, specific experiments can be designed to narrow down the algorithm. Behaviour and neuroimaging studies can provide evidence for and against these theories. Note also that current work in

neuroscience can shape these theories. For example, as early visual cells preferentially respond to specific orientations, algorithms about visual function need to incorporate orientation in models of early visual processing.

It is debatable at which of Marr's levels neuroimaging studies in Cognitive Neuroscience are describing function. The human neuroimaging studies using fMRI are recording from millions of neurons at the same time, and can therefore not describe the exact underlying hardware. In the example of human vision in the ventral temporal cortex, MVPA studies are exploring the representational dynamics between visual areas (Figure 1). Thus, MVPA is well suited to explore the algorithmic level, for example by mapping the areas involved in specific function, such as was done in chapter four of this thesis. However, as de-Wit also argued, the focus of this research needs to shift from mapping the information to showing how information is communicated within the brain (de-Wit et al., 2016). High temporal resolution neuroimaging data (e.g., MEG & EEG) has great potential to uncover dynamic transfer of information. For example, the temporal generalization method can be used to test theories about how information is formed and transferred over time (King et al., 2014; King and Dehaene, 2014). A strong focus on the development of methods that specifically test the read out and dynamics of information in the brain is critical for understanding its workings.

7 Conclusions

This thesis demonstrated that relating decoding methods to behaviour is important for drawing conclusions about how information is used by the brain. First, I expanded on the development of decoding methods, facilitating their application for time-series neuroimaging data. Secondly, I have built on the distance to bound approach as a linking method for brain decoding and behaviour. By testing specific predictions of this approach, I showed a systematic relationship between decoding methods and models of behaviour. Finally, I showed how the distance to bound approach can be used to distinguish between decodable information and information that is suitable for use in behaviour. Taken together, these results highlight that going beyond brain decoding is necessary to come to a complete understanding of cognitive function.

Whether neuroimaging can succeed in linking brain and cognition is heavily debated (Coltheart, 2006; de-Wit et al., 2016; Jonas and Kording, 2017; Krakauer et al., 2017; Poldrack, 2006, 2010). Some of these authors argue for a shift in perspective; rather than showing the availability of information, one must show how information is used by the brain (de-Wit et al., 2016; Klein, 2016; Naselaris and Kay, 2015; Ritchie et al., in press). Developing new ways of linking neuroimaging and behaviour is therefore a fundamental challenge for Cognitive Neuroscience. However, modelling the read out in behaviour is not sufficient, and methods for showing read out between different brain areas are also needed. The behavioural read out problem can be addressed by designing specific experiments to test approaches such as distance to bound for linking neuroimaging to behaviour. In the case of read out between brain areas, future work can develop new paradigms for testing information processing within the brain.

In conclusion, this thesis has argued that decodable information has to be interpreted with caution, as it does not imply that the information is being used by the brain. In addition, it has shown that by conducting a distance to bound analysis, evidence can be obtained supporting the use of information (decoded from neuroimaging data) in behaviour. Continuing to improve the methods for analysing neuroimaging data is an important requirement for the success of Cognitive Neuroscience. Developing new ways of investigating the dynamics of information within and between brain areas will be critical for a complete understanding of cognitive functions at all levels.

References

Allefeld, C., Görgen, K., Haynes, J.-D., 2016. Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. NeuroImage 141, 378–392. doi:10.1016/j.neuroimage.2016.07.040

Baillet, S., 2017. Magnetoencephalography for brain electrophysiology and imaging. Nat. Neurosci. 20, 327–339. doi:10.1038/nn.4504

Baillet, S., Friston, K.J., Oostenveld, R., 2011. Academic software applications for electromagnetic brain mapping using MEG and EEG. Comput. Intell. Neurosci. 2011, 972050.

Bennett, D., Murawski, C., Bode, S., 2015. Single-Trial Event-Related Potential Correlates of Belief Updating. eNeuro 2, ENEURO.0076-15.2015. doi:10.1523/ENEURO.0076-15.2015

Bi, Y., Wang, X., Caramazza, A., 2016. Object domain and modality in the ventral visual pathway. Trends Cogn. Sci. 20, 282–290.

Bracci, S., op de Beeck, H., 2016. Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J. Neurosci. 36, 432–444. doi:10.1523/JNEUROSCI.2314-15.2016

Brown, S.D., Heathcote, A., 2008. The simplest complete model of choice response time: Linear ballistic accumulation. Cognit. Psychol. 57, 153–178.

Bruner, J.S., Potter, M.C., 1964. Interference in visual recognition. Science 144, 424–425.

Caramazza, A., Mahon, B.Z., 2003. The organization of conceptual knowledge: the evidence from category-specific semantic deficits. Trends Cogn. Sci. 7, 354–361. doi:10.1016/S1364-6613(03)00159-1

Caramazza, A., Shelton, J.R., 1998. Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. J. Cogn. Neurosci. 10, 1–34. doi:10.1162/089892998563752

Carlson, T.A., 2014. Orientation Decoding in Human Visual Cortex: New Insights from an Unbiased Perspective. J. Neurosci. 34, 8373–8383. doi:10.1523/JNEUROSCI.0548-14.2014

Carlson, T.A., Ritchie, J.B., Kriegeskorte, N., Durvasula, S., Ma, J., 2014. Reaction Time for Object Categorization Is Predicted by Representational Distance. J. Cogn. Neurosci. 26, 132–142. doi:10.1162/jocn_a_00476

Carlson, T.A., Tovar, D.A., Alink, A., Kriegeskorte, N., 2013. Representational dynamics of object vision: The first 1000 ms. J. Vis. 13, 1. doi:10.1167/13.10.1

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Front. Neurosci. 6, 149.

Chan, A.M., Halgren, E., Marinkovic, K., Cash, S.S., 2010. Decoding word and category-specific spatiotemporal representations from MEG and EEG. NeuroImage 54, 3028–3039. doi:10.1016/j.neuroimage.2010.10.073

Cichy, R.M., Pantazis, D., Oliva, A., 2016. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. Cereb. Cortex N. Y. N 1991 26, 3563–3579. doi:10.1093/cercor/bhw135

Cichy, R.M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time. Nat. Neurosci. 17, 455–462. doi:10.1038/nn.3635

Coggan, D.D., Liu, W., Baker, D.H., Andrews, T.J., 2016. Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. NeuroImage 135, 107–114. doi:10.1016/j.neuroimage.2016.04.060

Coltheart, M., 2006. What has functional neuroimaging told us about the mind (so far)?(position paper presented to the european cognitive neuropsychology workshop, bressanone, 2005). Cortex 42, 323–331.

Contini, E.W., Wardle, S.G., Carlson, T.A., in press. Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. Neuropsychologia. doi:10.1016/j.neuropsychologia.2017.02.013

Coutanche, M.N., Solomon, S.H., Thompson-Schill, S.L., 2016. A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes. Neuropsychologia 82, 134–141. doi:10.1016/j.neuropsychologia.2016.01.018

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270. doi:10.1016/S1053-8119(03)00049-1

Crouzet, S.M., Kirchner, H., Thorpe, S.J., 2010. Fast saccades toward faces: Face detection in just 100 ms. J. Vis. 10, 16. doi:10.1167/10.4.16

Crouzet, S.M., Thorpe, S.J., 2011. Low-Level Cues and Ultra-Fast Face Detection. Front. Psychol. 2. doi:10.3389/fpsyg.2011.00342

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. 16, 763–770. doi:10.1038/nn.3381

de-Wit, L., Alexander, D., Ekroll, V., Wagemans, J., 2016. Is neuroimaging measuring information in the brain? Psychon. Bull. Rev. 23, 1415–1428. doi:10.3758/s13423-016-1002-0

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134, 9–21. doi:10.1016/j.jneumeth.2003.10.009

DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends Cogn. Sci. 11, 333–341. doi:10.1016/j.tics.2007.06.010

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 113, 7900–7905. doi:10.1073/pnas.1602413113

Forstmann, B.U., Wagenmakers, E.-J., 2015. An introduction to model-based cognitive neuroscience. Springer.

Gaspar, C.M., Rousselet, G.A., 2009. How do amplitude spectra influence rapid animal detection? Vision Res. 49, 3001–3012. doi:10.1016/j.visres.2009.09.021

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. Brain Imaging Methods 7, 267. doi:10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.S., 2014. MNE software for processing MEG and EEG data. NeuroImage 86, 446–460. doi:10.1016/j.neuroimage.2013.10.027

Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. 15, 536–548. doi:10.1038/nrn3747

Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., Pollmann, S., 2009a. PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. Neuroinformatics 7, 37–53. doi:10.1007/s12021-008-9041-y

Hanke, M., Halchenko, Y.O., Sederberg, P.B., Olivetti, E., Fründ, I., Rieger, J.W., Herrmann, C.S., Haxby, J.V., Hanson, S.J., Pollmann, S., Hanke, M., Halchenko, Y.O., Sederberg, P.B., Olivetti, E., Fründ, I., Rieger, J.W.,

Herrmann, C.S., Haxby, J.V., Hanson, S.J., Pollmann, S., 2009b. PyMVPA: a unifying approach to the analysis of neuroscientific data. Front. Neuroinformatics 3, 3. doi:10.3389/neuro.11.003.2009

Harel, A., Kravitz, D.J., Baker, C.I., 2014. Task context impacts visual object processing differentially across the cortex. Proc. Natl. Acad. Sci. 111, E962–E971. doi:10.1073/pnas.1312567111

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110. doi:10.1016/j.neuroimage.2013.10.067

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science 293, 2425–2430. doi:10.1126/science.1063736

Haynes, J.-D., 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87, 257–270. doi:10.1016/j.neuron.2015.05.025

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading Hidden Intentions in the Human Brain. Curr. Biol. 17, 323–328. doi:10.1016/j.cub.2006.11.072

Hebart, M.N., Görgen, K., Haynes, J.-D., 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinformatics 8, 88. doi:10.3389/fninf.2014.00088

IPCC, 2013. Climate change 2013: the physical science basis. Working group I contribution to the fifth assessment report of the intergovernmental panel on

climate change. Cambridge University Press, Cambridge, UK and New York, USA.

Jonas, E., Kording, K.P., 2017. Could a Neuroscientist Understand a Microprocessor? PLOS Comput. Biol. 13, e1005268. doi:10.1371/journal.pcbi.1005268

Kaiser, D., Azzalini, D.C., Peelen, M.V., 2016a. Shape-independent object category responses revealed by MEG and fMRI decoding. J. Neurophysiol. 115, 2246–2250. doi:10.1152/jn.01074.2015

Kaiser, D., Oosterhof, N.N., Peelen, M.V., 2016b. The Neural Dynamics of Attentional Selection in Natural Scenes. J. Neurosci. 36, 10522–10528. doi:10.1523/JNEUROSCI.1385-16.2016

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. J. Neurosci. 17, 4302–4311.

Kay, K.N., Weiner, K.S., Grill-Spector, K., 2015. Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. Curr. Biol. 25, 595–600. doi:10.1016/j.cub.2014.12.050

Khan, S.S., Madden, M.G., 2014. One-class classification: taxonomy of study and review of techniques. Knowl. Eng. Rev. 29, 345–374.

King, J.-R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. 18, 203–210. doi:10.1016/j.tics.2014.01.002

King, J.-R., Gramfort, A., Schurger, A., Naccache, L., Dehaene, S., 2014. Two Distinct Dynamic Modes Subtend the Detection of Unexpected Sounds. PLoS ONE 9, e85791. doi:10.1371/journal.pone.0085791

Klein, C., 2016. Brain regions as difference-makers. Philos. Psychol. 0, 1–14. doi:10.1080/09515089.2016.1253053

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490. doi:10.1016/j.neuron.2016.12.041

Kravitz, D.J., Saleem, K.S., Baker, C.I., Ungerleider, L.G., Mishkin, M., 2013. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn. Sci. 17, 26–49. doi:10.1016/j.tics.2012.10.011

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. Neuron 60, 1126–1141. doi:10.1016/j.neuron.2008.10.043

Long, B., Konkle, T., Cohen, M.A., Alvarez, G.A., 2016. Mid-level perceptual features distinguish objects of different real-world sizes. J. Exp. Psychol. Gen. 145, 95.

Luck, S.J., 2005. An Introduction to the Event-Related Potential Technique. MIT press.

Macé, M.J.-M., Delorme, A., Richard, G., Fabre-Thorpe, M., 2009. Spotting animals in natural scenes: efficiency of humans and monkeys at very low contrasts. Anim. Cogn. 13, 405–418. doi:10.1007/s10071-009-0290-4

Mack, M.L., Gauthier, I., Sadr, J., Palmeri, T.J., 2008. Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. Psychon. Bull. Rev. 15, 28–35. doi:10.3758/PBR.15.1.28

Mahon, B.Z., Anzellotti, S., Schwarzbach, J., Zampini, M., Caramazza, A., 2009. Category-Specific Organization in the Human Brain Does Not Require Visual Experience. Neuron 63, 397–405. doi:10.1016/j.neuron.2009.07.012

Mahon, B.Z., Milleville, S.C., Negri, G.A., Rumiati, R.I., Caramazza, A., Martin, A., 2007. Action-related properties shape object representations in the ventral stream. Neuron 55, 507–520.

Marr, D., 1982. Vision: A computational approach. MIT Press.

Maxfield, J.T., Stalder, W.D., Zelinsky, G.J., 2014. Effects of target typicality on categorical search. J. Vis. 14, 1–1. doi:10.1167/14.12.1

Minter, T.C., 1975. Single-class classification.

Mohan, K., Arun, S.P., 2012. Similarity relations in visual search predict rapid visual categorization. J. Vis. 12, 19–19. doi:10.1167/12.11.19

Naselaris, T., Kay, K.N., 2015. Resolving Ambiguities of MVPA Using Explicit Models of Representation. Trends Cogn. Sci. 19, 551–554. doi:10.1016/j.tics.2015.07.005

Nastase, S.A., Connolly, A.C., Oosterhof, N.N., Halchenko, Y.O., Guntupalli, J.S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M.I., Haxby, J.V., 2017. Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. Cereb. Cortex 1–15. doi:10.1093/cercor/bhx138

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430. doi:10.1016/j.tics.2006.07.005

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2010. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput. Intell. Neurosci. 2011, 156869. doi:10.1155/2011/156869

Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front. Neuroinformatics 27. doi:10.3389/fninf.2016.00027

Peelen, M.V., Bracci, S., Lu, X., He, C., Caramazza, A., Bi, Y., 2013. Tool selectivity in left occipitotemporal cortex develops without vision. J. Cogn. Neurosci. 25, 1225–1234.

Peelen, M.V., He, C., Han, Z., Caramazza, A., Bi, Y., 2014. Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. J. Neurosci. 34, 163–170.

Philiastides, M.G., Sajda, P., 2006. Temporal Characterization of the Neural Correlates of Perceptual Decision Making in the Human Brain. Cereb. Cortex 16, 509–518. doi:10.1093/cercor/bhi130

Poldrack, R.A., 2010. Mapping mental function to brain structure: how can cognitive neuroimaging succeed? Perspect. Psychol. Sci. 5, 753–761.

Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10, 59–63. doi:10.1016/j.tics.2005.12.004

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126. doi:10.1038/nrn.2016.167

Proklova, D., Kaiser, D., Peelen, M.V., 2016. Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. J. Cogn. Neurosci. 1–13. doi:10.1162/jocn_a_00924

Purcell, B.A., Heitz, R.P., Cohen, J.Y., Schall, J.D., Logan, G.D., Palmeri, T.J., 2010. Neurally Constrained Modeling of Perceptual Decision Making. Psychol. Rev. 117, 1113–1143. doi:10.1037/a0020311

Purcell, B.A., Palmeri, T.J., 2016. Relating accumulator model parameters and neural dynamics. J. Math. Psychol. doi:10.1016/j.jmp.2016.07.001

Rae, B., Heathcote, A., Donkin, C., Averell, L., Brown, S., 2014. The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. J. Exp. Psychol. Learn. Mem. Cogn. 40, 1226–1243. doi:10.1037/a0036801

Ricciardi, E., Bonino, D., Pellegrini, S., Pietrini, P., 2014. Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture? Neurosci. Biobehav. Rev. 41, 64–77.

Ritchie, J.B., Bracci, S., op de Beeck, H., in press. Avoiding illusory effects in representational similarity analysis: what (not) to do with the diagonal. NeuroImage. doi:10.1016/j.neuroimage.2016.12.079

Ritchie, J.B., Carlson, T.A., 2016. Neural Decoding and “Inner” Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. Front. Neurosci. 190. doi:10.3389/fnins.2016.00190

Ritchie, J.B., Kaplan, D.M., Klein, C., in press. Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. Br. J. Philos. Sci.

Ritchie, J.B., Tovar, D.A., Carlson, T.A., 2015. Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. PLoS Comput Biol 11, e1004316. doi:10.1371/journal.pcbi.1004316

Ritter, C., Hebart, M.N., Wolbers, T., Bingel, U., 2014. Representation of Spatial Information in Key Areas of the Descending Pain Modulatory System. J. Neurosci. 34, 4634–4639. doi:10.1523/JNEUROSCI.4342-13.2014

Rogers, T.T., Hocking, J., Mechelli, A., Patterson, K., Price, C., 2005. Fusiform Activation to Animals is Driven by the Process, Not the Stimulus. J. Cogn. Neurosci. 17, 434–445. doi:10.1162/0898929053279531

Rosch, E.H., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P., 1976. Basic objects in natural categories. Cognit. Psychol. 8, 382–439.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366.

Strother, S.C., 2006. Evaluating fMRI preprocessing pipelines. IEEE Eng. Med. Biol. Mag. 25, 27–41. doi:10.1109/MEMB.2006.1607667

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. 2011, 8.

Tarr, M.J., Gauthier, I., 2000. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nat. Neurosci. 3, 764–769.

Thirion, B., Pedregosa, F., Eickenberg, M., Varoquaux, G., 2015. Correlations of correlations are not reliable statistics: implications for multivariate pattern analysis., in: ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015).

Van Veen, B.D., Van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44, 867–880. doi:10.1109/10.623056

Varoquaux, G., Raamana, P., Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2016. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. ArXiv160605201 Stat.

Varoquaux, G., Thirion, B., 2014. How machine learning is shaping cognitive neuroimaging. GigaScience 3, 28. doi:10.1186/2047-217X-3-28

Wang, X., He, C., Peelen, M.V., Zhong, S., Gong, G., Caramazza, A., Bi, Y., 2017. Domain selectivity in the parahippocampal gyrus is predicted by the same structural connectivity patterns in blind and sighted individuals. J. Neurosci. 3622–16. doi:10.1523/JNEUROSCI.3622-16.2017

Wardle, S.G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., Carlson, T.A., 2016. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. NeuroImage 132, 59–70. doi:10.1016/j.neuroimage.2016.02.019

Wardle, S.G., Ritchie, J.B., Seymour, K., Carlson, T.A., 2017. Edge-Related Activity Is Not Necessary to Explain Orientation Decoding in Human Visual Cortex. J. Neurosci. 37, 1187–1196. doi:10.1523/JNEUROSCI.2690-16.2016

Williams, M.A., Dang, S., Kanwisher, N.G., 2007. Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. 10, 685–686. doi:10.1038/nn1900

Wu, C.-T., Crouzet, S.M., Thorpe, S.J., Fabre-Thorpe, M., 2015. At 120 msec You Can Spot the Animal but You Don't Yet Know It's a Dog. J. Cogn. Neurosci. 27, 141–149. doi:10.1162/jocn_a_00701

Zelinsky, G.J., Adeli, H., Peng, Y., Samaras, D., 2013. Modelling eye movements in a categorical search task. Phil Trans R Soc B 368, 20130058. doi:10.1098/rstb.2013.0058

Appendix A of this thesis has been removed as it may contain sensitive/confidential content